STATISTICAL METHODS COMMONLY USED IN WATER QUALITY DATA ANALYSIS

DEVELOPED BY

STANLEY L. PONCE

WSDG TECHNICAL PAPER
WSDG-TP-00001
DECEMBER 1980

WATERSHED SYSTEMS DEVELOPMENT GROUP
USDA FOREST SERVICE
3825 EAST MULBERRY STREET
FORT COLLINS, COLORADO 80524

Table of Contents

Table of Contents

List of Tables

List of Tables

List of Tables

List of Figures

List of Examples

PREFACE

The purpose of this WSDG Technical Paper is to review several statistical methods commonly used in water quality data analysis. It is not intended to be another detailed statistical text, but rather, provide Forest Hydrologists with a concise review of basic statistics. Throughout the paper the assumptions underlying the various statistical tests have been emphasized. In addition, water quality examples have been included throughout. Although it is anticipated that you will perform most of your statistical analyses on programmable calculators or computers, the basic formula and procedures involved in the various statistical tests have been included to help give you a better perspective of what is going on in the "canned" programs.

This Technical Paper was designed to be used in conjunction with WSDG Technical Paper 00002, "Water Quality Monitoring Programs," and WSDG Application Documents 00001 and 00002, Statistical Analysis Using SAS at the USEPA National Computer Center" and "Statistical Analysis Using SPSS at the USDA Fort Collins Computer Center," respectively. Together, these two Technical Papers and two Support Documents form the "Water Quality Handbook" scheduled for release late winter of 1981 by the USDA Forest Service.

I would like to acknowledge all the people that contributed to the development of this Technical Paper. A very special thanks is given to Dave Ryn, Computer Specialist and Statistician, WSDG; Walt Hivner, Statistician, Colorado State University, and Bob Averett, Hydrologist, USGS-WRD, who performed the detailed technical review. I would like to extend my appreciation to Jim Ingwersen, Cindi Eichin, Gordon Snyder, Eric

Siverts, and all the Forest Hydrologists who provided many helpful suggestions concerning content and format.

STATISTICAL METHODS COMMONLY USED IN WATER QUALITY DATA ANALYSIS

1.0 Introduction

The proper use of statistical methods in the reduction and analysis of water quality data is very important. These methods allow us to take an objective view of the data and, hopefully, keep us from making fools of ourselves by claiming that a favorite theory is substantiated by observations that do nothing of the sort (Colquhoun, 1971). This is not to imply that field observations are not important in the interpretive process, for indeed they are. However, these observations should be tempered with the objectivity of statistical methods. Averett (1979) states

> "data interpretation is an intellectual activity; statistical application is a mechanical activity. Good experimental design eases the interpretation burden, and statistical methods are best applied in evaluating the adequacy or applicability of the selected experimental design. Statistical methods are, therefore, extremely useful tools in the reduction and analysis of data, and as such, can be used as an aid in data interpretation. The statistical testing of data can provide 'yes' or 'no' answers at a given probability level--nothing more."

The purpose of this WSDG Technical Paper is to review several selected statistical methods. It is not intended to be another detailed statistical text. If you feel you need more information concerning a specific statistical method, two texts recommended are Biometry by Sokal and Rohlf (1969) and Statistical Analysis of Samples of Benthic Invertebrates by Elliott (1977).

This paper begins with a series of definitions of common statistical terms followed by (1) a brief discussion of the underlying assumptions and data requirements for parametric statistical testing, (2) descriptive statistics, (3) important theoretical probability density functions, (4) confidence limits about the mean, (5) hypothesis testing, and (6) testing the homogeneity of the variance. It is important that you understand these concepts before proceeding, for they are fundamental to an understanding of the subsequent statistical methods. Next, the powerful tools of (1) analysis of variance and (2) regression and correlation are covered. Throughout this paper, emphasis has been placed on the assumptions underlying the various methods and the utility of these methods for the analysis of water quality data.

To enhance your understanding of the various statistical methods presented, water quality-related examples have been included throughout. It is realized that you will be doing most of your data analysis using statistical packages designed for hand-held calculators or computers. However, it is important that you clearly understand what the "canned" programs are doing. Therefore, I have outlined the basic "number crunching" operations associated with each method. The statistical package that you use for your data analysis is generally a function of availability. Two packages which are readily available to Forest Hydrologists are SAS (Statistical Analysis System, 1979) and SPSS (Statistical Package for Social Sciences, 1975). You can address SAS through the EPA NCC-IBM (where STORET data is retained), and SPSS through the USDA Fort Collins Computer Center. Many of the examples have been solved using SAS and SPSS. A detailed discussion of the procedures necessary to address the various statistical methods using SAS and SPSS is presented by Ingwersen (1981a and 1981b).

2.0 Basic Definitions

It seems that any discussion of statistics always begins with a series of definitions. It helps to get everyone into the same framework for statistical thinking and, as such, is necessary. What follows is a brief journey through some of the basic definitions fundamental to the science of statistics. Only the ones necessary to get us on our way are presented here. As we proceed into the maze of statistical methods and concepts, this list of definitions will grow.

In a narrow sense, statistics refers to any of the various computed or estimated statistical quantities, such as the mean, standard deviation, or variance. In a broad sense, it can be thought of as the scientific study of numerical data from natural phenomena, concerned with making inferences about a population based upon information on a sample taken from the population.

A collection of individual observations obtained by a specified procedure, such as random sampling, is commonly referred to as a sample or a sample of observations. If the sample has been collected properly, it can be considered to be representative of a population containing all possible observations. The specific property measured by the individual observations is called the variable. Variables are subdivided into several categories, two of which concern us. Continuous variables, such as calcium or total dissolved solids concentrations, theoretically can assume an infinite number of values between any two fixed points and are limited only by the precision of the instrument used to make the measurement. Discontinuous or discrete variables, such as fecal coliform or stonefly nymph counts, are those that have only certain fixed numerical values with no intermediate values possible inbetween. Usually, discrete data take on integral numerical values, such as 1, 2, 3, 100, or 200.

3.0 Underlying Assumptions and Data Requirements for Parametric Statistical Testing

Most water quality data analysis performed by Forest Hydrologists will employ parametric statistical methods. These methods are the ones commonly taught at the undergraduate level in college and are characterized by using tests whose models specify conditions about the parameters of the population from which the sample was drawn. There are several basic assumptions underlying parametric statistical tests of which you should be aware whenever you apply these procedures. These are:

1. the observations are independent;
2. the distribution of the population is known;
3. the variances of the populations being compared are equal or of known ratio.

In addition, these statistical tests require that the data:

1. were collected in a random manner;
2. have error variation independent of the means, normally distributed and homogeneous; and
3. have variance components which are additive.

The data commonly encountered in water quality studies rarely satisfy all the assumptions and requirements of parametric statistical methods. However, as Glass and others (1972) state, the relevant question is not whether the assumptions and requirements are met exactly, but rather whether the plausible violations of the assumptions and requirements have serious consequences on the validity of the probability statements based upon them. At this point, it is only important that you keep these assumptions and requirements in mind. As we develop our statistical arsenal, I will illustrate procedures which can be applied to the raw data to test if the requirements are met and to point out the pitfalls you may encounter in your inferences when these assumptions and requirements are violated.

4.0 Descriptive Statistics

Data analysis generally begins with numerical summary or description of the data set. Two types of descriptive statistics are addressed here: statistics of central tendency and statistics of dispersion.

The statistics of central tendency or location include the mean, median, and mode. The mean is the one we are most familiar with and is a statistic of great importance since many statistical tests center around comparison of the sample means. The arithmetic mean ($\bar{X}$) is calculated by summing all the individual observations (ΣX_i) of a sample and dividing it by the number of observations (n) in the sample (Equation 1).

$$\bar{X} = \frac{\Sigma X_i}{n} \qquad (1)$$

As you recall, sample statistics are designed to be estimators of population parameters. In this case $\bar{X}$ is an estimate of μ, the population mean. (Note: As a matter of convention in statistics, Greek letters are used to denote population parameters while Roman letters are used to denote sample statistics.)

The median divides a frequency distribution into two halves and is defined as that value of a variable (in an array ordered from the smallest to the largest) that has an equal number of observations on either side of it. The median is easily determined when the sample array has an odd number of observations. However, when the number of observations is even, it is determined by calculating the mean of the (n/2)th and [(n/2) + 1]th observations. In certain cases, when the distribution is asymmetric, as is generally the case with coliform data, the median is a more representative measure of location than the arithmetic mean.

The mode is defined as the value or class interval having the greatest number of individuals. In some cases, a mode will not exist, such as when several values or frequency classes have the same number of observations.

Table 1 summarizes the characteristics, advantages and disadvantages of the arithmetic mean, median and mode. In symmetrical distributions, such as the normal distribution, the mean, median and mode are all identical. In asymmetrical distributions the mean is relatively closer to the drawn-out tail of the distribution, while the mode is farthest from the mean and the median lies between the two (Figure 1).

The statistics of dispersion include the range, variance and standard deviation. The range is simply the difference between the largest and smallest observations in a sample. Since the range can be affected greatly by an extreme value, it should be considered only as a rough estimate of the dispersion of all the observations in a sample.

Both the variance and standard deviation take all observations of a sample into consideration and weigh each observation by its distance (deviation) from the calculated center (mean) of the distribution. The deviation (x_i) of the ith observation (X_i) from the mean ($\bar{X}$) can be expressed as

$$x_i = X_i - \bar{X}$$

The variance (s^2) is a modified average of the sum of the deviations squared (ΣX_i^2) (Equation 2). The variance is expressed in squared units.

$$s^2 = \frac{\Sigma x_i^2}{n-1} \qquad (2)$$

Table 1. A summary of the characteristics, advantages and disadvantages of the arithmetic mean, median and mode.

	MEAN	MEDIAN	MODE
CHARACTERISTICS	1. It is a calculated average. 2. It is affected greatly by extreme values. 3. The sum of the deviations about the arithmetic mean is zero. 4. The sum of the squares of the deviations from the arithmetic mean is less than those computed about any other point. 5. The sum of the means equals the mean of the sums. 6. It has a smaller standard error than other statistics of location.	1. It is a measure of central tendency based on position. 2. It is affected by the number of items, not by the size of the extreme values. 3. The sum of the absolute value of the deviations about the median will be less than any other point.	1. It is a measure of central tendency based on position. 2. It is the most typical value. 3. It is entirely independent of extreme values.
ADVANTAGES	1. It is the most commonly used, easily understood and generally recognized average. 2. It is simple to calculate and uses all observations in the sample. 3. It may be treated algebraically.	1. It is not distorted by extreme values, and therefore, may be more representative of central tendency than other measures.	1. It is considered to be the most descriptive measure of central tendency since it is the most typical value.
DISADVANTAGES	1. It may be greatly distorted by extreme values and therefore may not be a typical value.	1. It is not as commonly used nor understood as the arithmetic mean. 2. The observations must be ranked by magnitude before the median can be computed. 3. It has a larger standard error than the arithmetic mean. 4. It cannot be manipulated algebraically.	1. If none of the values occur more than once, the mode does not exist. 2. A sample may have more than one mode.

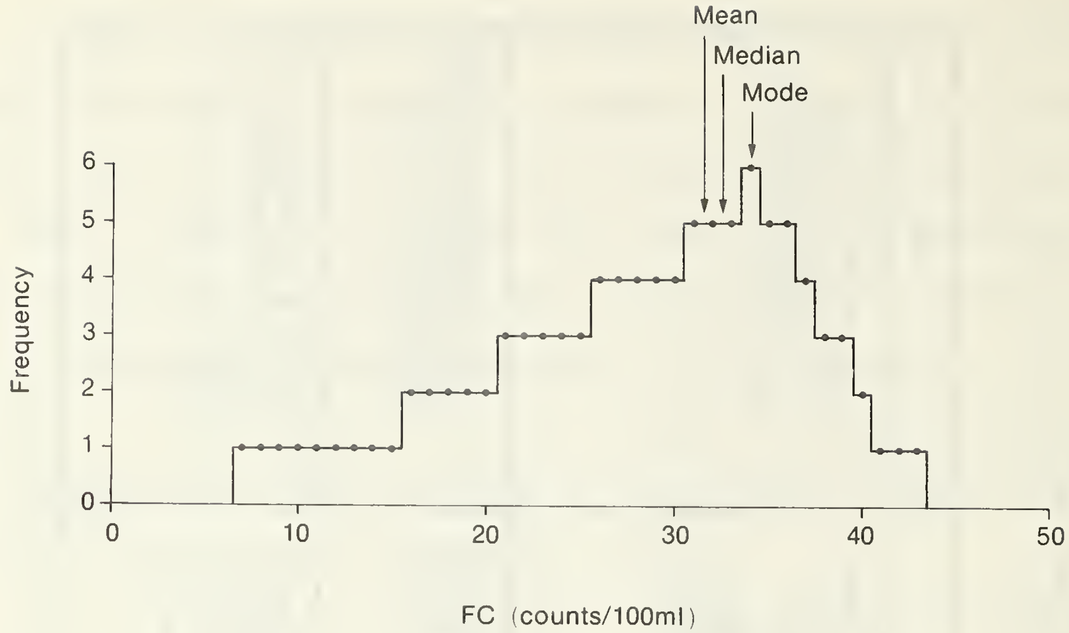


Figure 1. Relative position of the mean, median and mode for an asymmetrical distribution.

The sum of the deviations is divided by "n-1" in order to obtain an unbiased estimate of the population variance. Statisticians have shown that when the sums of squares (the deviations squared) are divided by "n-1" rather than "n" a more accurate estimate of the population variance will be obtained, no matter what the sample size. Calculations that divide the sum of squares by "n" tend to underestimate the variance, especially when "n" is small, and result in a biased estimate of the population variance. Consequently, you should always use "n-1" in your calculation of the estimate of the population variance. The quantity "n-1" is generally referred to as the degrees of freedom (df). As you will see later, the degrees of freedom are used with many other statistics besides the variance.

The standard deviation (s) is calculated by taking the square root of the variance (Equation 3).

$$s = \sqrt{\frac{\Sigma x_i^2}{n-1}} \tag{3}$$

The advantage of the standard deviation (s) over the variance (s^2) is that the units are not squared and, therefore, tend to be more meaningful.

Sometimes we wish to compare the amount of variation in populations having different means. To accomplish this we use the coefficient of variation (CV) which is simply the standard deviation expressed as a percentage of the mean (Equation 4).

$$CV = \frac{s(100)}{\bar{X}} \tag{4}$$

A summary of the characteristics, advantages and disadvantages of the range, standard deviation and coefficient of variation is presented in Table 2.

Use of the descriptive statistics described in this section is illustrated in Examples 1a and 1b. The examples utilize a population of specific conductance values and fecal coliform counts, ranked in order of increasing magnitude (Table 3).

Table 2. A summary of the characteristics, advantages and disadvantages of the range, standard deviation and coefficient of variation.

	RANGE	STANDARD DEVIATION	COEFFICIENT OF VARIATION
CHARACTERISTICS	1. It is dependent only on the two extremes; i.e. the highest and lowest observations.	1. It is a measure of the dispersion of observations based on their distribution.	1. It is a relative measure of variation.
ADVANTAGES	1. Simple and easily understood. 2. Readily determined. 3. No information about the distribution is required to determine the range.	1. Simple and easy to determine. 2. It is affected by the value of every observation. 3. It is commonly used and understood.	1. Simple and easy to determine. 2. It is independent of the unit of measurement used.
DISADVANTAGES	1. It yields no information about how the items are dispersed.	1. Standard deviations of constituents with different units, such as ppm and counts/100 ml, cannot be readily compared.	1. To know whether or not a particular value of the coefficient of variation is unusually large or small requires experience with similar data.

Table 3. Populations of specific conductance (SC) and fecal coliform (FC), ranked in order of increasing magnitude, with means (μ) of 81.5 μmhos/cm and 23.6 counts/100 ml and standard deviations (σ) of 12.1 μmhos/cm and 12.5 counts/100 ml, respectively.

Rank	SC	FC	Rank	SC	FC	Rank	SC	FC	Rank	SC	FC	Rank	SC	FC
1	51	7	11	67	12	21	72	15	31	75	16	41	79	18
2	56	8	12	67	13	22	72	15	32	75	17	42	79	19
3	58	9	13	68	13	23	73	15	33	76	17	43	79	19
4	59	10	14	68	13	24	73	15	34	77	17	44	79	19
5	61	10	15	68	13	25	74	15	35	77	17	45	80	19
6	62	11	16	69	14	26	74	16	36	78	17	46	80	19
7	63	11	17	69	14	27	74	16	37	78	18	47	80	20
8	64	11	18	70	14	28	74	16	38	78	18	48	81	20
9	65	12	19	71	14	29	75	16	39	78	18	49	81	20
10	66	12	20	72	14	30	75	16	40	78	18	50	82	20

Rank	SC	FC	Rank	SC	FC	Rank	SC	FC	Rank	SC	FC	Rank	SC	FC
51	82	21	61	84	23	71	88	26	81	92	29	91	98	40
52	82	21	62	85	23	72	88	26	82	92	30	92	99	44
53	82	21	63	85	24	73	88	27	83	93	30	93	100	45
54	82	21	64	86	24	74	89	27	84	93	31	94	100	46
55	83	22	65	86	24	75	89	27	85	94	31	95	101	49
56	83	22	66	86	24	76	89	28	86	94	32	96	103	54
57	83	22	67	86	25	77	90	28	87	95	32	97	105	59
58	83	22	68	87	25	78	90	28	88	95	33	98	107	61
59	83	23	69	87	25	79	91	29	89	96	37	99	108	67
60	84	23	70	87	26	80	92	29	90	97	38	100	111	75

EXAMPLE 1a
Descriptive Statistics

Problem:

Obtain two random samples, containing 15 observations each, from the specific conductance and fecal coliform populations tabulated in Table 3. For each sample determine the (a) mean, (b) standard deviation, and (c) coefficient of variation.

Solution:

The samples were obtained with the aid of a random number (RN) table.

RN	SC	FC	RN	SC	FC	RN	SC	FC
59	83	23	38	78	18	24	73	15
34	77	17	9	65	12	27	74	16
46	80	19	95	101	49	65	86	24
81	92	29	60	84	23	48	81	20
55	83	22	78	90	28	35	77	17

(a) Mean

$$\overline{SC} = \Sigma SC_i/n = 1224/15 = 81.6\ \mu\text{mhos/cm}$$

$$\overline{FC} = \Sigma FC_i/n = 331/15 = 22.1\ \text{counts/100ml}$$

(b) Standard Deviation

$$s_{(SC)} = \sqrt{\frac{\Sigma(SC_i - \overline{SC})^2}{n-1}} = 8.7\ \mu\text{mhos/cm}$$

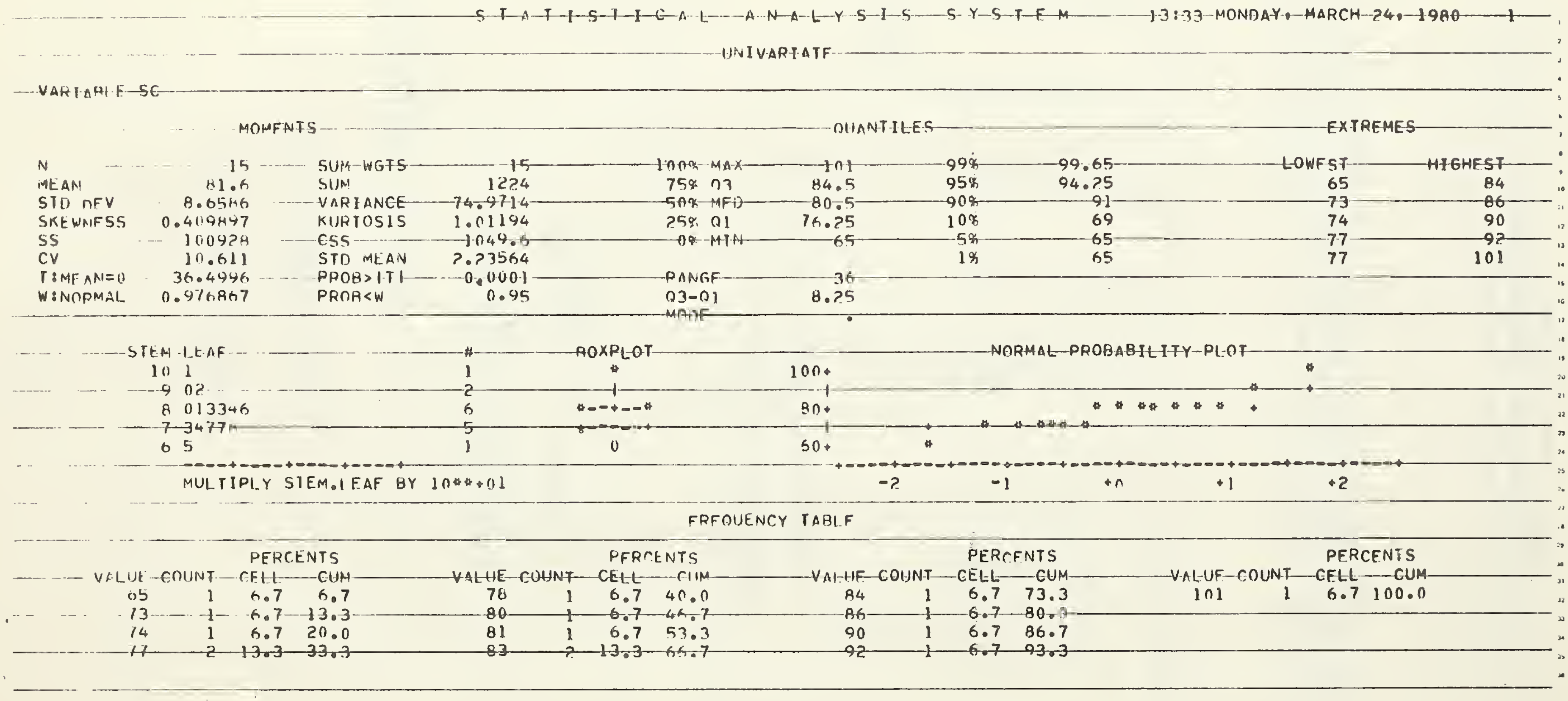
$$s_{(FC)} = \sqrt{\frac{\Sigma(FC_i - \overline{FC})^2}{n-1}} = 8.8\ \text{counts/100ml}$$

(c) Coefficient of Variation

$$CV_{(SC)} = \frac{s_{SC} \times 100}{\overline{SC}} = 10.6\%$$

$$CV_{(FC)} = \frac{s_{FC} \times 100}{\overline{FC}} = 39.8\%$$

The UNIVARIATE procedure from SAS (1979) or the CONDESCRIPTIVE subprogram from SPSS (1975) could be used to solve this example problem. The UNIVARIATE procedure was used and the results are presented in Tables 4 and 5.



```
                              S T A T I S T I C A L   A N A L Y S I S   S Y S T E M        13:33 MONDAY, MARCH 24, 1980    1

                                                         UNIVARIATE

VARIABLE=SC

                  MOMENTS                                                         QUANTILES                                           EXTREMES

N                15     SUM WGTS          15          100% MAX       101       99%       99.65                  LOWEST       HIGHEST
MEAN           81.6     SUM             1224           75% Q3       84.5       95%       94.25                      65            84
STD DEV      8.65866    VARIANCE     74.9714           50% MED      80.5       90%          91                      73            86
SKEWNESS    0.409897    KURTOSIS     1.01194           25% Q1      76.25       10%          69                      74            90
SS            100928    CSS           1049.6            0% MIN        65        5%          65                      77            92
CV            10.611    STD MEAN     2.23564                                   1%          65                      77           101
T:MEAN=0     36.4996    PROB>|T|      0.0001          RANGE           36
W:NORMAL    0.976867    PROB<W          0.95          Q3-Q1         8.25
                                                      MODE             .

         STEM LEAF                       #          BOXPLOT                               NORMAL PROBABILITY PLOT
           10 1                          1             *                100+                                              *
            9 02                         2             |                   |                                        *     +
            8 013346                     6          *--+--*              80+                        * *  ** * * *     +
            7 34778                      5          +-----+                |            +     *  *  ***  *
            6 5                          1             0                 50+            *
              ----+----+----+----+                                          +----+----+----+----+----+----+----+----+----+----+
         MULTIPLY STEM.LEAF BY 10**+01                                           -2        -1        +0        +1        +2

                                                      FREQUENCY TABLE

                PERCENTS                            PERCENTS                            PERCENTS                            PERCENTS
   VALUE COUNT  CELL   CUM         VALUE COUNT  CELL   CUM         VALUE COUNT  CELL   CUM         VALUE COUNT  CELL   CUM
      65     1   6.7   6.7            78     1   6.7  40.0            84     1   6.7  73.3           101     1   6.7 100.0
      73     1   6.7  13.3            80     1   6.7  46.7            86     1   6.7  80.0
      74     1   6.7  20.0            81     1   6.7  53.3            90     1   6.7  86.7
      77     2  13.3  33.3            83     2  13.3  66.7            92     1   6.7  93.3
```

Table 4. SAS output for the specific conductance portion of Part 2, Example 1a.

```
                                        S T A T I S T I C A L   A N A L Y S I S   S Y S T E M          13133 MONDAY, MARCH 24, 1980       2

                                                              UNIVARIATE

 VARIABLE FC

                  MOMENTS                                                    QUANTILES                                         EXTREMES

 N                    15    SUM WGTS          15          100% MAX        49        99%          46               LOWEST       HIGHEST
 MEAN            22.1333    SUM              332           75% Q3      23.25        95%          34                   12            23
 STD DEV         8.79827    VARIANCE     77.4095           50% MED      19.5        90%        28.5                   15            24
 SKEWNESS        2.16738    KURTOSIS     6.14815           25% Q1      16.75        10%        13.5                   16            28
 SS                 8432    CSS          1083.73            0% MIN        12         5%          12                   17            29
 CV              39.7512    STD MEAN      2.2717                                     1%          12                   17            49
 T:MEAN=0        9.74306    PROB>|T|      0.0001          RANGE           37
 W:NORMAL       0.796151    PROB<W          0.01          Q3-Q1          6.5
                                                          MODE             .

            STEM LEAF                         #      BOXPLOT                                         NORMAL PROBABILITY PLOT
               4 9                            1         *                  40+                                              *
               3                                                             |                                              +
               2 0233489                      7      +--+--+                 |                       ++ + **** *  *  *
               1 2567789                      7      *-----*               10+          *     *  * *** ** *
                 ----+----+----+----+                                         +----+----+----+----+----+----+----+----+----+----+
                 MULTIPLY STEM.LEAF BY 10**+01                                     -2        -1        +0        +1        +2

                                                               FREQUENCY TABLE

                      PERCENTS                           PERCENTS                            PERCENTS                            PERCENTS
         VALUE COUNT  CELL   CUM        VALUE COUNT  CELL   CUM        VALUE COUNT  CELL   CUM        VALUE COUNT  CELL   CUM
            12     1   6.7   6.7           18     1   6.7  40.0           23     2  13.3  73.3           49     1   6.7 100.0
            15     1   6.7  13.3           19     1   6.7  46.7           24     1   6.7  80.0
            16     1   6.7  20.0           20     1   6.7  53.3           28     1   6.7  86.7
            17     2  13.3  33.3           22     1   6.7  60.0           29     1   6.7  93.3
```

Table 5. SAS output for the fecal coliform portion of Part 2, Example 1a.

EXAMPLE 1b
Descriptive Statistics

Problem:

For the specific conductance and fecal coliform populations tabulated in Table 3 determine the (a) median, (b) mode, and (c) range.

Solution:

(a) Median

The median for the specific conductance population is the mean of ranks 50 and 51, which is 82 μmhos/cm.

The median for the fecal coliform population is 20.5 counts/100 ml.

(b) Mode

The specific conductance population does not have a mode; 82 and 83 are both observed five times.

The mode of the fecal coliform population is 16 counts/100 ml.

(c) Range

Specific conductance: 111 - 51 = 60 μmhos/cm
Fecal coliform: 75 - 7 = 68 counts/100 ml.

The UNIVARIATE procedure from SAS (1979) was also used to develop the required descriptive statistics (Tables 6 and 7). Note that there is no data entry for the mode of the specific conductance since it does not exist. The CONDESCRIPTIVE subprogram from SPSS (1975) could also have been used, however, it will not determine the median and mode.

```
                                       S T A T I S T I C A L   A N A L Y S I S   S Y S T E M        14:55 MONDAY, MARCH 24, 1980    1

                                                             UNIVARIATE

VARIABLE SC

                  MOMENTS                                              QUANTILES                                   EXTREMES

N                100     SUM WGTS         100          100% MAX        111       99%          108              LOWEST       HIGHEST
MEAN           81.53     SUM             8153          75% Q3           89       95%          101                  51           103
STD DEV      12.0818     VARIANCE     145.969          50% MED          82       90%           97                  56           105
SKEWNESS   0.0417951     KURTOSIS   -0.120319          25% Q1           74       10%           66                  58           107
SS            679165     CSS          14450.9          0% MIN           51        5%           61                  59           108
CV           14.8188     STD MEAN     1.20818                                     1%           51                  61           111
T:MEAN=0     67.4819     PROB>|T|      0.0001          RANGE            60
D:NORMAL     0.42124     PROB>D             1          Q3-Q1            15
                                                       MODE              .

     STEM LEAF                               #      BOXPLOT                          NORMAL PROBABILITY PLOT
       11 1                                  1         0             110+                                                  *
       10 0013578                            7         0                |                                         **** *   +
        9 0012223344556789                  16         |                |                                 ********++++ +
        8 00011222223333344556666777888999  32      *--+--*          80+                        *********++++
        7 012223344445555677888889999       27      +-----+             |           +++++*********
        6 1234567788899                     13         |                |+   + ++*********
        5 1689                               4         0             50+*   * **
          ----+----+----+----+----+----+                                +----+----+----+----+----+----+----+----+----+----+
          MULTIPLY STEM.LEAF BY 10**+01                                      -2        -1        +0        +1        +2

                                                          FREQUENCY TABLE

                   PERCENTS                        PERCENTS                        PERCENTS                        PERCENTS
     VALUE COUNT  CELL    CUM        VALUE COUNT  CELL    CUM        VALUE COUNT  CELL    CUM        VALUE COUNT  CELL    CUM
        51     1   1.0    1.0           70     1   1.0   18.0           83     5   5.0   59.0           96     1   1.0   89.0
        56     1   1.0    2.0           71     1   1.0   19.0           84     2   2.0   61.0           97     1   1.0   90.0
        58     1   1.0    3.0           72     3   3.0   22.0           85     2   2.0   63.0           98     1   1.0   91.0
        59     1   1.0    4.0           73     2   2.0   24.0           86     4   4.0   67.0           99     1   1.0   92.0
        61     1   1.0    5.0           74     4   4.0   28.0           87     3   3.0   70.0          100     2   2.0   94.0
        62     1   1.0    6.0           75     4   4.0   32.0           88     3   3.0   73.0          101     1   1.0   95.0
        63     1   1.0    7.0           76     1   1.0   33.0           89     3   3.0   76.0          103     1   1.0   96.0
        64     1   1.0    8.0           77     2   2.0   35.0           90     2   2.0   78.0          105     1   1.0   97.0
        65     1   1.0    9.0           78     5   5.0   40.0           91     1   1.0   79.0          107     1   1.0   98.0
        66     1   1.0   10.0           79     4   4.0   44.0           92     3   3.0   82.0          108     1   1.0   99.0
        67     2   2.0   12.0           80     3   3.0   47.0           93     2   2.0   84.0          111     1   1.0  100.0
        68     3   3.0   15.0           81     2   2.0   49.0           94     2   2.0   86.0
        69     2   2.0   17.0           82     5   5.0   54.0           95     2   2.0   88.0
```

Table 6. SAS output for the specific conductance portion of Part 2, Example 1b.

```
                           S T A T I S T I C A L   A N A L Y S I S   S Y S T E M         14:55 MONDAY, MARCH 24, 1980     2

                                                      UNIVARIATE

VARIABLE FC

                  MOMENTS                                            QUANTILES                                  EXTREMES

N               100   SUM WGTS        100       100% MAX   75      99%   66.9999              LOWEST       HIGHEST
MEAN          23.55   SUM            2355        75% Q3    27      95%        49                   7            54
STD DEV      12.546   VARIANCE    157.402        50% MED   20      90%        38                   8            59
SKEWNESS    1.87556   KURTOSIS     4.1614        25% Q1    15      10%        12                   9            61
SS            71043   CSS         15582.8         0% MIN    7       5%        10                  10            67
CV          53.2738   STD MEAN     1.2546                           1%         7                  10            75
T:MEAN=0     18.771   PROB>|T|     0.0001       RANGE      68
D:NORMAL    1.47602   PROB>D    0.0256269       Q3-Q1      12
                                                MODE       16

   STEM LEAF                                      #   BOXPLOT                        NORMAL PROBABILITY PLOT
      7 5                                         1      *          70+                                                  *
      6 17                                        2      *            |                                          *  *
      5 49                                        2      *            |                                        **
      4 04569                                     5      *            |                                     ***      +
      3 001122378                                 9      0            |                                 +****+++++  +
      2 00001111222233334444555666777888999      35   +-----+         |                        +++++*********
      1 0011122233333444445555556666666777778888899 43 *-----*        |              ***************
      0 789                                       3      |           0+*     * ***
        ----+----+----+----+----+----+----+----+                      +----+----+----+----+----+----+----+----+----+----+
        MULTIPLY STEM.LEAF BY 10**+01                                      -2        -1        +0        +1        +2

                                                     FREQUENCY TABLE

              PERCENTS                        PERCENTS                         PERCENTS                         PERCENTS
VALUE COUNT  CELL   CUM   VALUE COUNT  CELL    CUM    VALUE COUNT  CELL    CUM    VALUE COUNT  CELL    CUM
    7     1   1.0   1.0      17     5   5.0   36.0       27     3   3.0   75.0       44     1   1.0   92.0
    8     1   1.0   2.0      18     5   5.0   41.0       28     3   3.0   78.0       45     1   1.0   93.0
    9     1   1.0   3.0      19     5   5.0   46.0       29     3   3.0   81.0       46     1   1.0   94.0
   10     2   2.0   5.0      20     4   4.0   50.0       30     2   2.0   83.0       49     1   1.0   95.0
   11     3   3.0   8.0      21     4   4.0   54.0       31     2   2.0   85.0       54     1   1.0   96.0
   12     3   3.0  11.0      22     4   4.0   58.0       32     2   2.0   87.0       59     1   1.0   97.0
   13     4   4.0  15.0      23     4   4.0   62.0       33     1   1.0   88.0       61     1   1.0   98.0
   14     5   5.0  20.0      24     4   4.0   66.0       37     1   1.0   89.0       67     1   1.0   99.0
   15     5   5.0  25.0      25     3   3.0   69.0       38     1   1.0   90.0       75     1   1.0  100.0
   16     6   6.0  31.0      26     3   3.0   72.0       40     1   1.0   91.0
```

Table 7. SAS output for the fecal coliform portion of Part 2, Example 1b.

5.0 Important Theoretical Probability Density Functions

Statisticians have developed several theoretical probability density functions which can be used as models for samples from a population. If the frequency distribution of observations from a sample fits one of these models, then (1) errors of the population parameters can be estimated, (2) temporal and spatial changes in frequency can be compared and (3) the effect of environmental factors and management practices can be examined.

The models commonly applied to discontinuous or count data are the *Binomial* and the *Poisson* distributions. A detailed discussion and illustration of the application of each of these models is given by Elliott (1977). The most important theoretical probability density functions for continuous or measured data are the *normal*, *t*, *chi-square* and *F* distributions. The balance of this Section will address these distributions. It is very important that you have a clear understanding of these distributions since many of the procedures that follow are based on a knowledge of these models.

5.1 Normal Distribution

5.1.1 Introduction

The normal distribution is also referred to as the "Gausian" distribution. We are all familiar with the symmetrical bell-shaped curve of this distribution. However, we should note that not all symmetrical distributions are normally distributed since the normal distribution is defined by a specific equation and has its own distinctive set of properties. The equation for the normal curve is:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{X-\mu}{\sigma}\right)^2} \qquad (5)$$

where Y is the height of the curve corresponding to an assigned value of X (which represents the frequency of observations of a given X_i); π and e are constants nearly equal to 3.1416 and 2.7183, respectively; and the other parameters are as previously defined. The properties of normal distributions are: (1) they are symmetrical about the mean (the mean, median and mode are all equal); (2) they all have a concave downward trend when $X < \pm \sigma$ of the mean and a concave upward trend when $X > \pm \sigma$ of the mean; and (3) the proportion of the area between two X values is completely governed by μ and σ.

Generally, we are interested in the area under the curve since we can relate it to the probability of occurrence of an interval of observations. (This concept of probability is generally denoted: Pr (argument) = area.) The area contained under the curve between two observations, such as X_a and X_b, can be determined by using Table A-1 (found in Appendix A).

In order that the same table can be used for any value of μ and σ, which vary with different normal populations, Equation 5 was standardized by substituting z for $\frac{X-\mu}{\sigma}$ prior to integration. Table A-1 is entered with the value of z which is simply the deviation of X from μ measured in σ units. The area between any two X_i's can be found by using the symmetry of the curve about z = 0. The standardized normal curve is illustrated in Figure 2. It is apparent that 68.3% of the observations lie within $\pm \sigma$ of the mean, 95.5% within $\pm 2\sigma$ of the mean and 99.7% within $\pm 3\sigma$ of the mean. Application of the normal distribution is illustrated in Example 2.

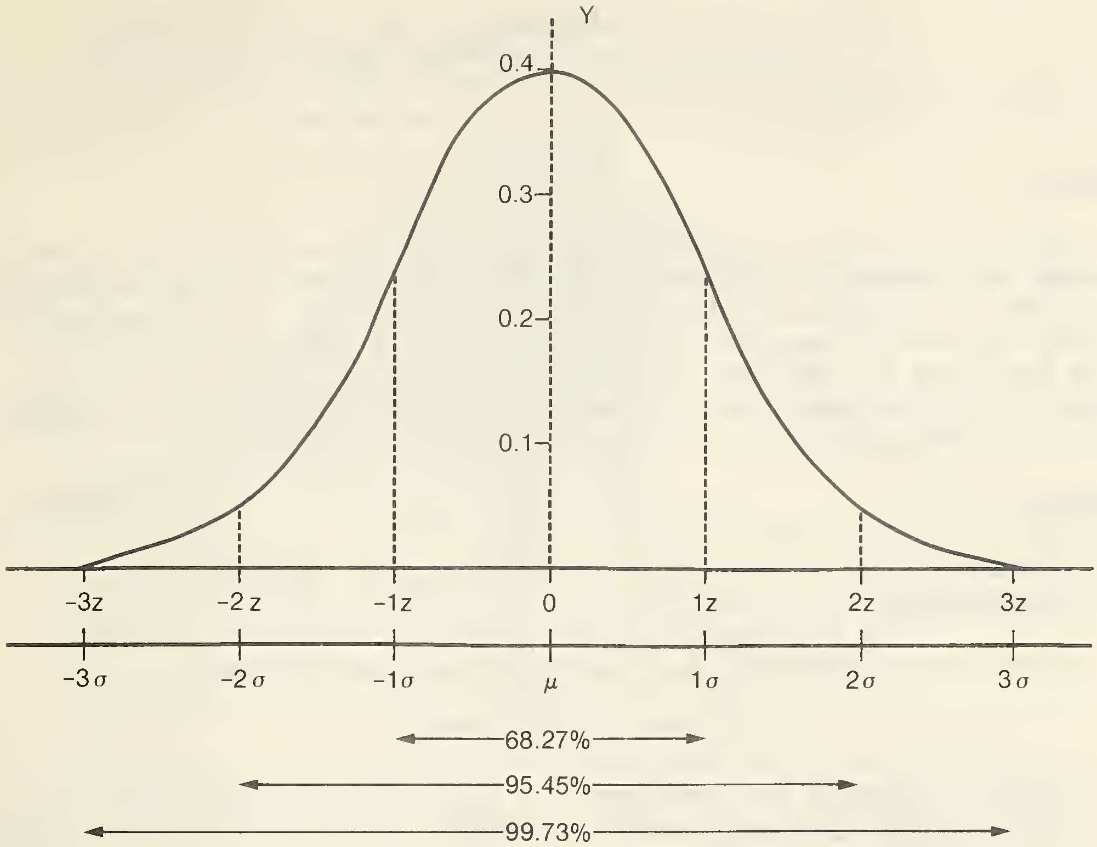


Figure 2. The standardized normal curve.

The relationship between z and σ is illustrated by the use of two ordinate scales.

EXAMPLE 2
Application of the Normal Curve

Problem:

For the specific conductance population presented in Table 3, which is assumed to be normally distributed with mean 81.5 µmhos/cm and standard deviation 12.1 µmhos/cm, determine (a) the probability of an observation falling between the mean and 91 µmhos/cm, (b) the probability of a value being less than 87 µmhos/cm, (c) the probability of a value occurring between 77 and 94 µmhos/cm, (d) the 90% value of X, and (e) the probability of X_i = 54 µmhos/cm.

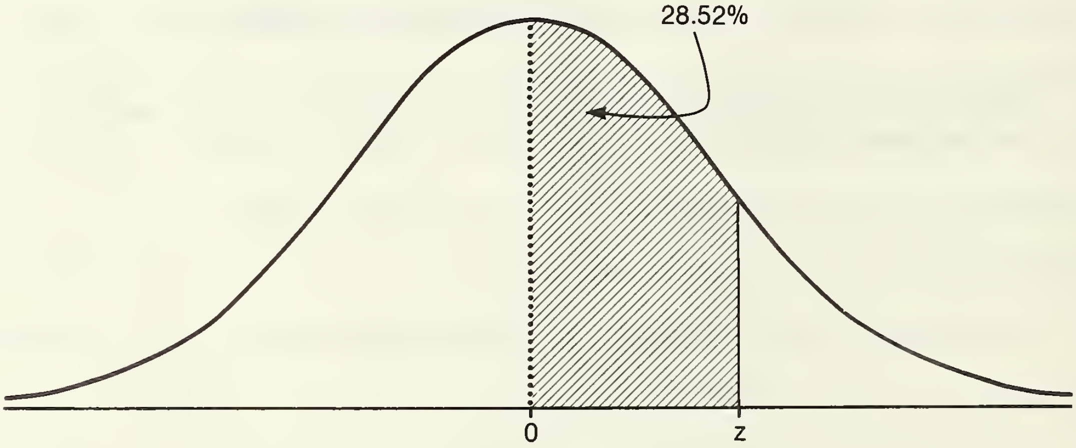
Solution:

(a) Probability of an observation falling between the mean and 91 µmhos/cm.

First, it is necessary to convert 91 µmhos/cm to z units.

$$z = \frac{X - \mu}{\sigma} = \frac{91 - 81.5}{12.1} = 0.79$$

Second, determine the area from z = 0 to z = 0.79 using Table A-1. Area = 0.2852 (0.5000 - 0.2148 = 0.2852). The area we are interested in is illustrated below.



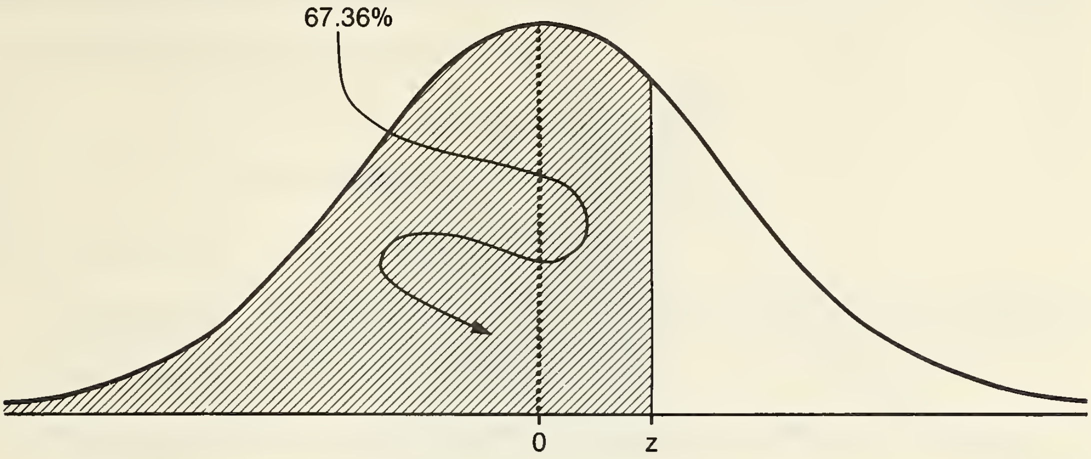
Therefore, the Pr $(81.5 \leq X_i \leq 91) = 0.2852$

(b) Probability of a value being less than 87 µmhos/cm.

First, convert 87 µmhos/cm to z units.

$$z = \frac{87 - 81.5}{12.1} = 0.45$$

Second, determine the area from Table A-1.
Area = 0.5 + 0.1736 =0.6736.
The area we are interested in is illustrated below.

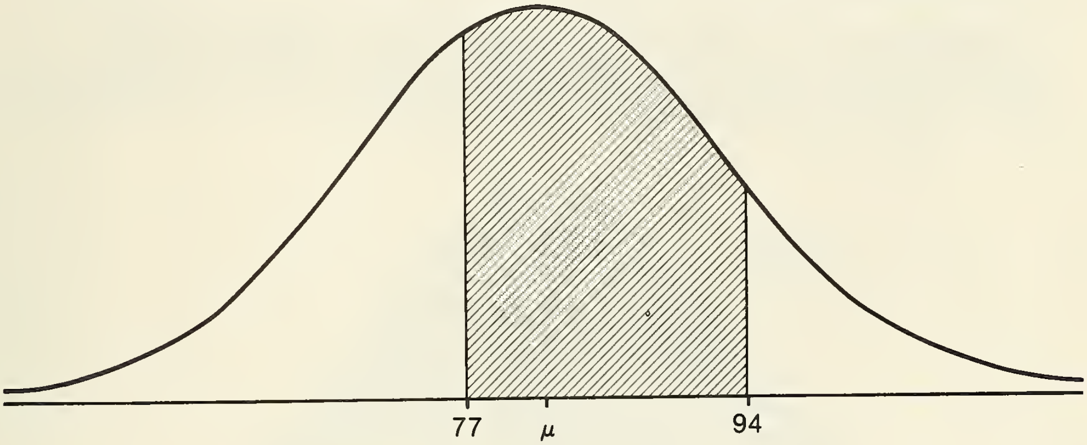


Since both sides of the curve are included in the argument, we must add 0.5000 to the value obtained from Table A-1.

Therefore, $\Pr(X_i \leq 87) = 0.6736$

(c) Probability of a value occurring between 77 and 94 μmhos/cm.

In this case, the argument falls on both sides of the mean.



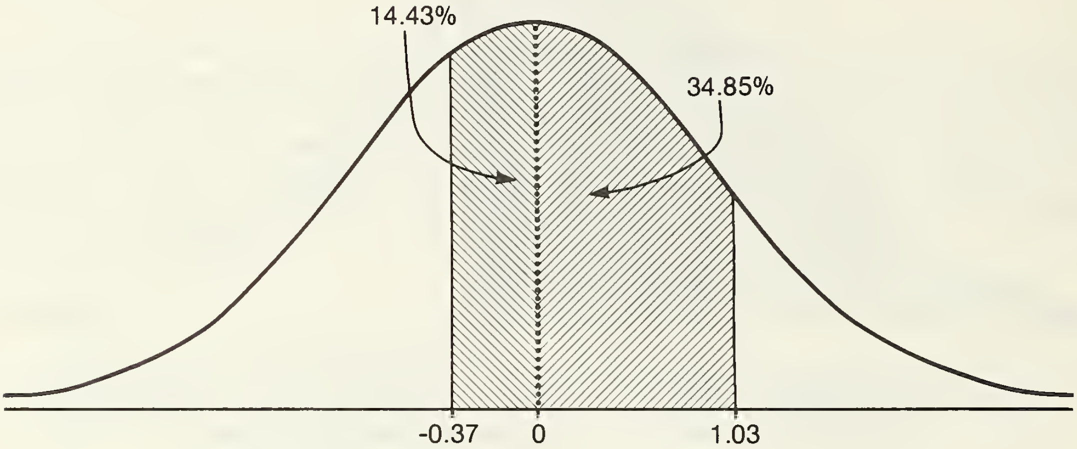
The area is calculated in three steps:

First, determine the z values.

$$z = \frac{94 - 81.5}{12.1} = 1.03$$

$$z = \frac{77 - 81.5}{12.1} = -0.37$$

Second, determine the area and apply the concept of symmetry to the curve (Table A-1). The area to the right of the center is equal to 0.3485 while the area to the left of the center is 0.1443 (see figure below).

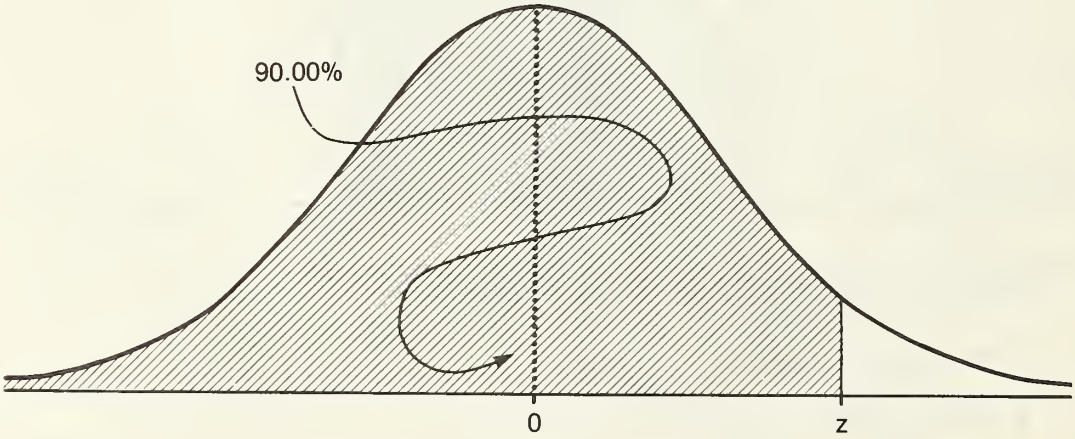


Third, add the two areas.

$\Pr (77 < X_i < 94) = 0.3485 + 0.1443 = 0.4928$

(d) The 90% value of X.

To determine the value of X_i for which 90% (0.9000) of the observations will fall below, you must first find the 0.4000 value in Table A-1, then determine z and solve for X_i.



The 0.4000 value was obtained by applying the concept of symmetry (0.4000 = 0.9000 - 0.5000).

From Table A-1 we find that z is nearly equal to 1.28.

Since: $z = (X - \mu)/\sigma$

It follows that:

$$X = 2\sigma + \mu$$

Substituting and solving we get:

$$X = 1.28\ (12.1) + 81.5$$

$$X = 97.0\ \mu\text{mhos/cm}$$

(e) Probability of X_i = 54 μmhos/cm.

To apply the normal distribution to discrete data it is necessary to treat the data as if it were continuous. Thus, the value 54 is considered as 53.5 to 54.5 μmhos/cm.

Now the problem becomes similar to that presented in item (c).

$Pr(53.5 \leq X \leq 54.5) = 0.0025$

5.1.2 Skewness and Kurtosis

Many times an observed frequency distribution will depart markedly from a normal distribution. There are two types of departures with which we need to be familiar: skewness and kurtosis.

Skewness refers to the asymmetry of the data where one tail is drawn out more than the other. A distribution skewed to the right has a long right tail. Kurtosis refers to the degree of peakedness of the curve. A distribution that has more observations near the mean and at the tails with fewer observations in the intermediate regions relative to the normal distribution with the same mean and variance is called leptokurtic. A platykurtic curve is the opposite of the leptokurtic curve and has more observations in the intermediate regions than at the mean or in the tails relative to the normal distribution.

In general, these statistics are seldom used in water quality data analysis. However, they are presented here simply as definitions for use in future discussions.

5.1.3 Testing the Normality of the Distribution

As you recall, one of the assumptions underlying parametric statistics is that the distribution of the population is known. Many water quality problems have statistical answers based on the assumption that the distribution of the population is normal. There are several methods you can use to check this assumption. Three procedures are presented here: (1) the graphic method, (2) the Kolmogorov-Smirnov test, and (3) the Shapiro-Wilk test. (It should be noted that most statistics books suggest the use of the chi-square test for goodness of fit. However, because the Kolmogorov-Smirnov and Shapiro-Wilk tests are more powerful and are the ones commonly contained in computer statistical packages, they are the ones addressed here to test for normality.)

The graphic method is based on a cumulative frequency distribution. When data from a normal distribution are plotted in a cumulative manner on arithmetic graph paper, the result is a sigmoid curve. This curve can be linearized by plotting the cumulative distribution on normal probability graph paper. Figure 3 illustrates a series of frequency distributions departing from normality. You will find these as useful guides when examining the distributions of your data on probability paper. The graphic method is illustrated in Example 3.

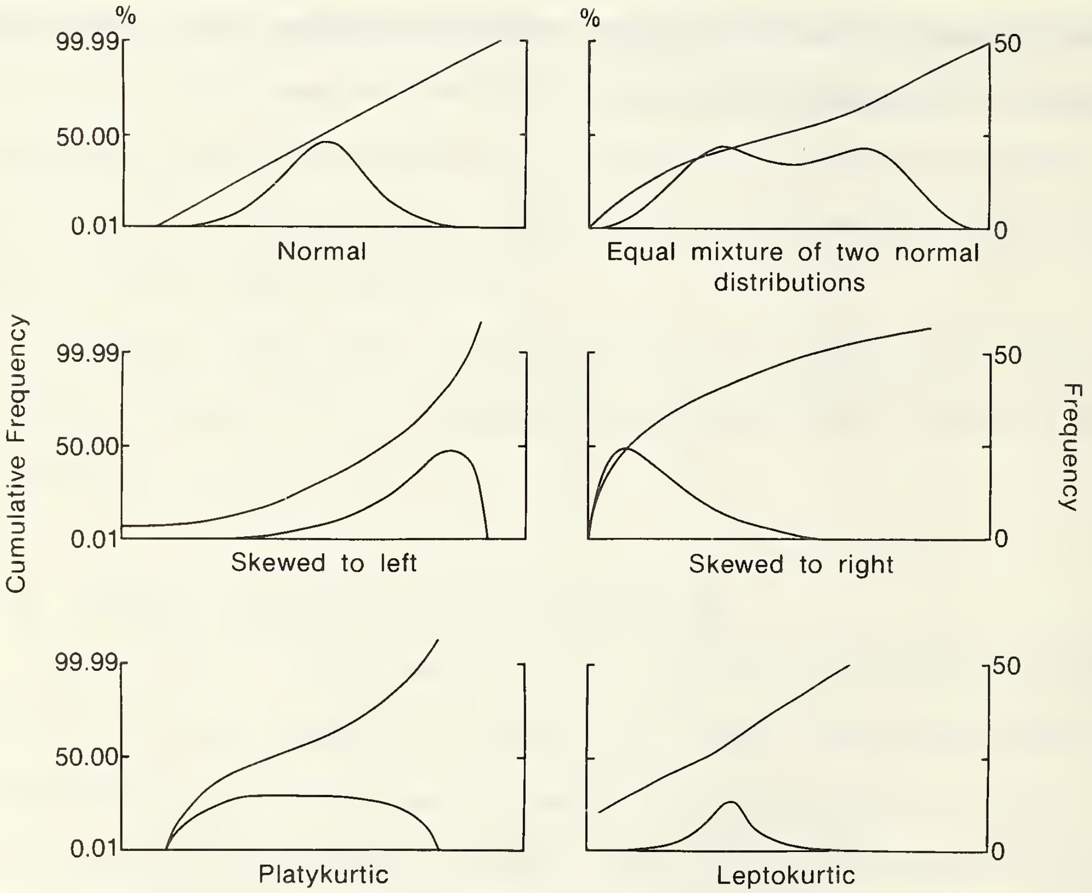


Figure 3. Examples of several frequency distributions and their respective cumulative frequency distributions (after Sokal and Rohlf, 1969).

EXAMPLE 3
Graphical Test for Normality

Problem:

Graphically test the populations presented in Table 3 for normality of their frequency distribution.

Solution:

(a) Prepare a frequency distribution and cumulative frequency distribution for (1) specific conductance and (2) fecal coliform.

SPECIFIC CONDUCTANCE

CLASS INTERVAL[1/] LOWER LIMIT (μmhos/cm)	UPPER LIMIT (μmhos/cm)	FREQUENCY (f)	CUMULATIVE FREQUENCY (F)	CUMULATIVE PERCENT FREQUENCY (% F) [2/]
50.5	55.5	1	1	1
55.5	60.5	3	4	4
60.5	65.5	5	9	9
65.5	70.5	9	18	18
70.5	75.5	14	32	32
75.5	80.5	15	47	47
80.5	85.5	16	63	63
85.5	90.5	15	78	78
90.5	95.5	10	88	88
95.5	100.5	6	94	94
100.5	105.5	3	97	97
105.5	110.5	2	99	99
110.5	115.5	1	100	100

[1/] Steel and Torrie (1960) suggest some guidelines for determining the size of the class interval: A rule of use in determining the size of the class interval when high precision is required in calculations made from the resulting frequency table is to make the interval not greater than one-quarter of the standard deviation. If this rule is strictly adhered to, the data are sometimes not sufficiently summarized for graphical presentation. If the size of the class interval is increased to one-third to one-half of a standard deviation, the resulting frequency table will usually be a sufficient summary for graphical presentation and adequate for most data; the lack of precision in any statistics calculated from the table will be small enough to be ignored.

[2/] Take care to note that since 100 observations are contained in the example population, the cumulative frequency (f) and cumulative percent frequency (% F) columns are identical. However, in the case where the cumulative frequency (F) is not equal to 100, the cumulative percent frequency (% F) is determined by dividing the observations in each class interval (f) by the total number of observations in the sample.

FECAL COLIFORM

CLASS INTERVAL				CUMULATIVE
LOWER	UPPER		CUMULATIVE	PERCENT
LIMIT	LIMIT	FREQUENCY	FREQUENCY	FREQUENCY
(counts/100 ml)		(%)	(F)	(% F)
6.5	10.5	5	5	5
10.5	14.5	15	20	20
14.5	18.5	21	41	41
18.5	22.5	17	58	58
22.5	26.5	14	72	72
26.5	30.5	11	83	83
30.5	34.5	5	88	88
34.5	38.5	2	90	90
38.5	42.5	1	91	91
42.5	46.5	3	94	94
46.5	50.5	1	95	95
50.5	54.5	1	96	96
54.5	58.5	0	96	96
58.5	62.5	2	98	98
62.5	66.5	0	98	98
66.5	70.5	1	99	99
70.5	74.5	0	99	99
74.5	78.5	1	100	100

(b) Graph the cumulative % F versus class interval (Figure 4) for specific conductance and fecal coliform on normal probability paper. Fit a straight line to each data set giving weight to those points occurring between cumulative frequencies of 25% to 75%.

(c) The specific conductance data follow the straight pretty well, strongly suggesting a normal distribution, while the trend of the fecal coliform population suggests that it is skewed to the right (Figure 4). A plot of the actual frequency distribution is included for your reference (Figure 5 and 6).

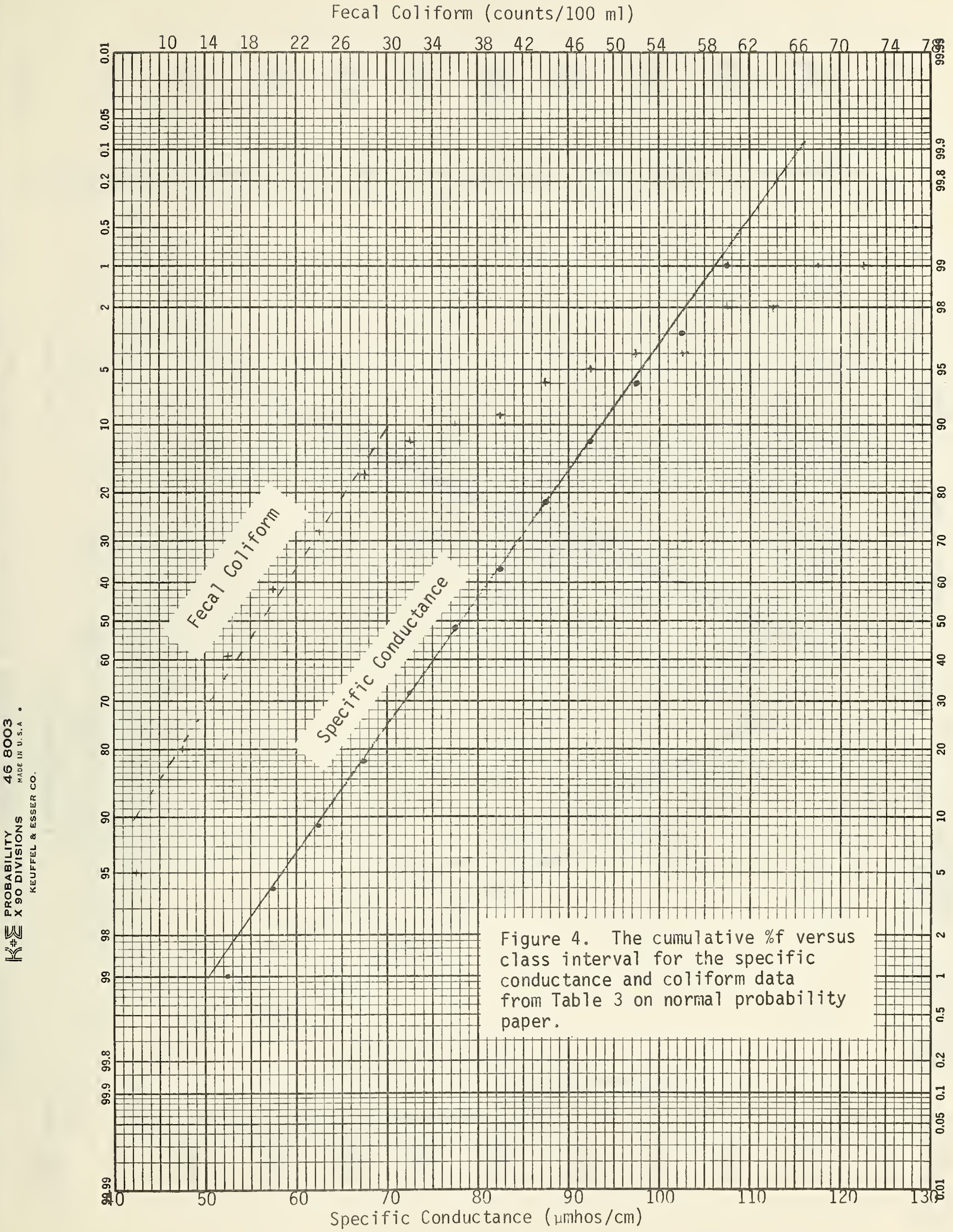


Figure 4. The cumulative %f versus class interval for the specific conductance and coliform data from Table 3 on normal probability paper.

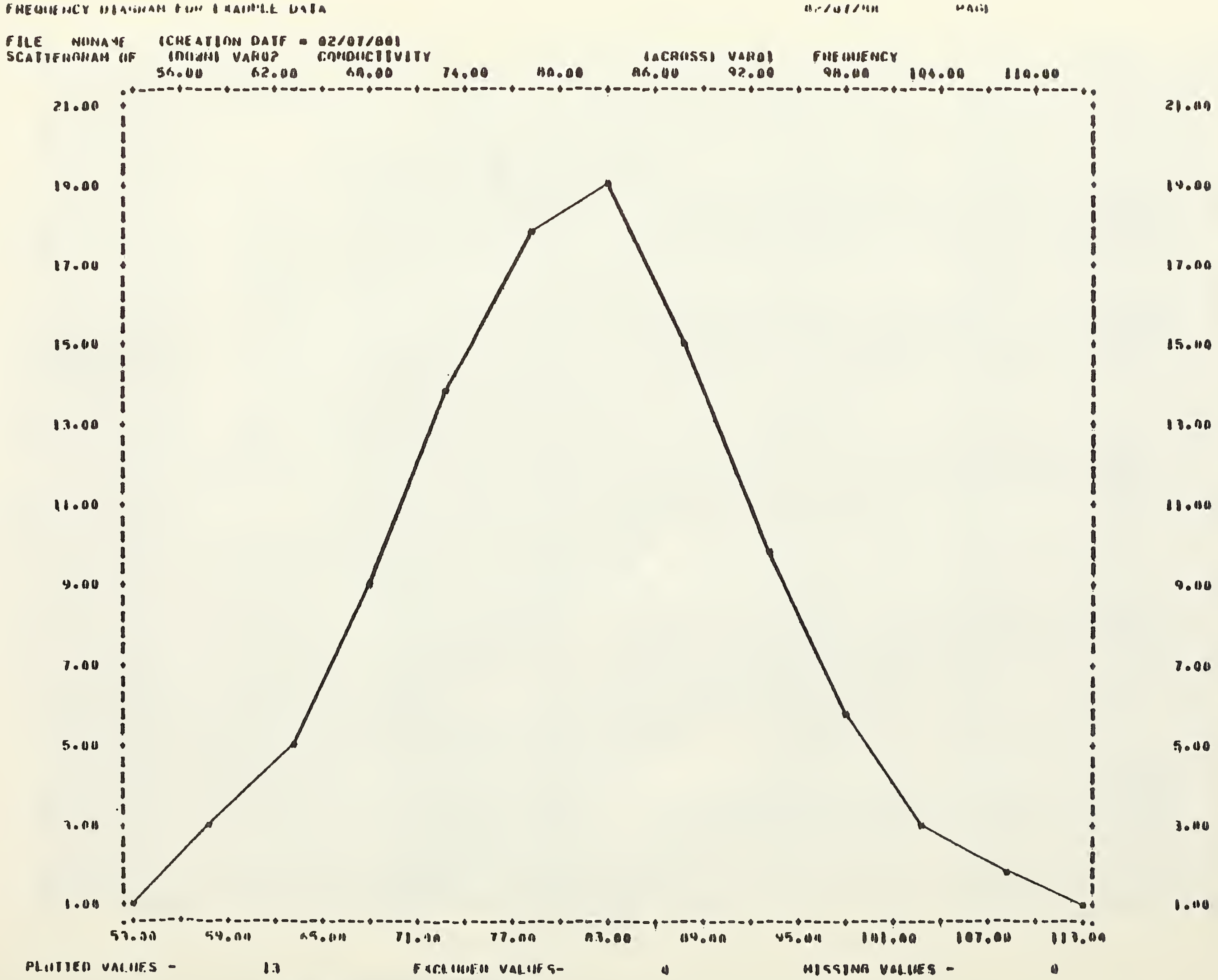


Figure 5. Frequency diagram for the conductivity data from Table 3.
Note that the lines connecting the plot points have been added to the SPSS plot.

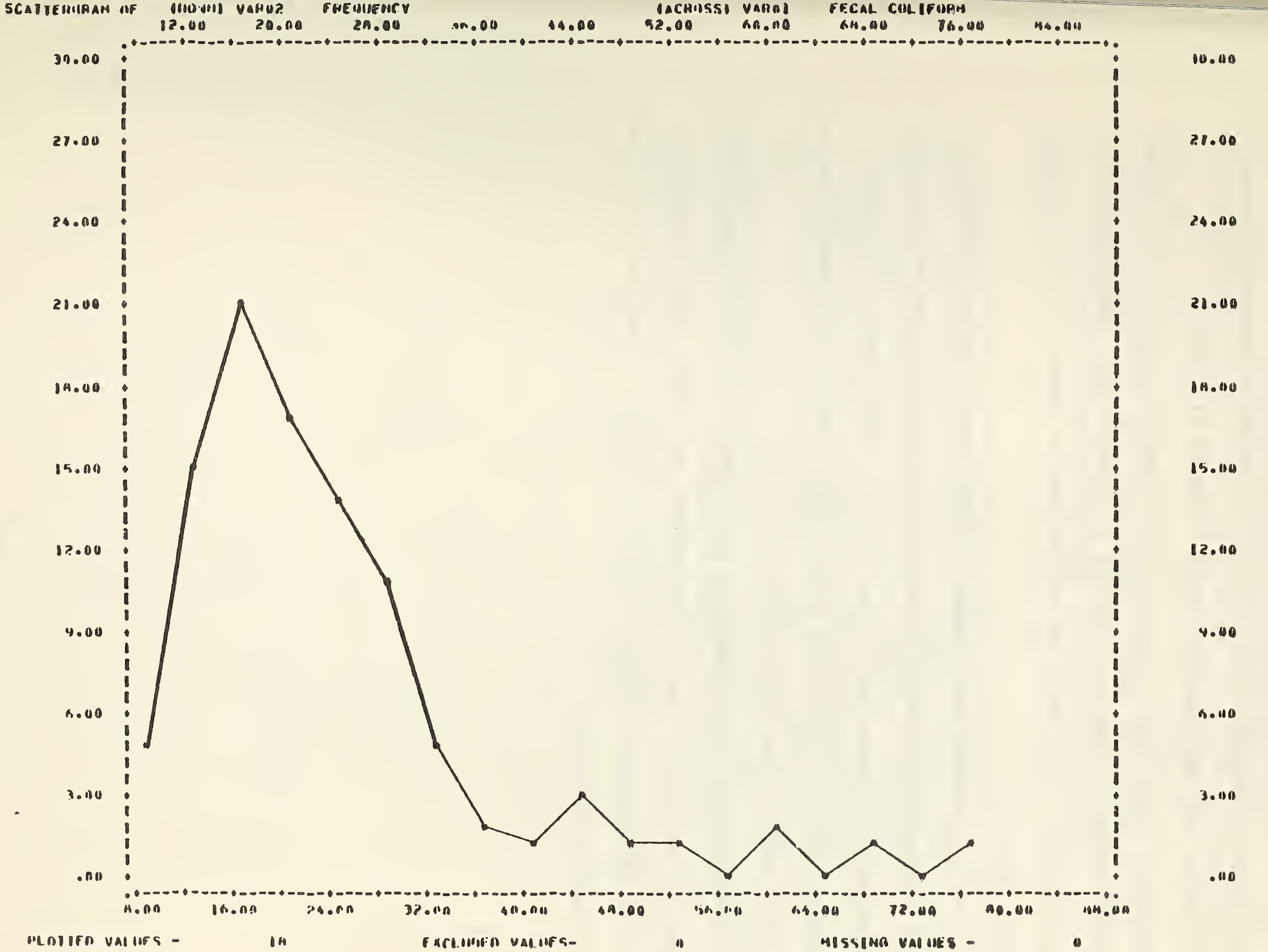


Figure 6. Frequency plot of the fecal coliform data from Table 3. Note that the lines connecting the plot points have been added to the SPSS plot.

The Shapiro-Wilk and Kolmogorov tests for goodness of fit are both computational procedures which allow us to compare an observed frequency distribution with the expected normal frequency distribution. The Shapiro-Wilk test (Shapiro and Wilk, 1965) should be used when the number of observations is less than or equal to 50 while the Kolmogorov-Smirnov test (Stephens, 1974) should be applied when the number of observations is greater than 50.

The Shapiro-Wilk test produces a W-statistic. The null hypothesis (no difference between the observed and the expected normal frequency distribution) is rejected (see Section 7.0) for small values of W. The Kolmogorov-Smirnov test yields a D-statistic. The null hypothesis, in this case, is rejected for large values of D.

Since both of these procedures are rather involved, the computational steps are not outlined here. Both methods are readily accessible through SAS and the Kolmogorov-Smirnov D-statistic in SPSS. These are illustrated in Examples 4a and 4b.

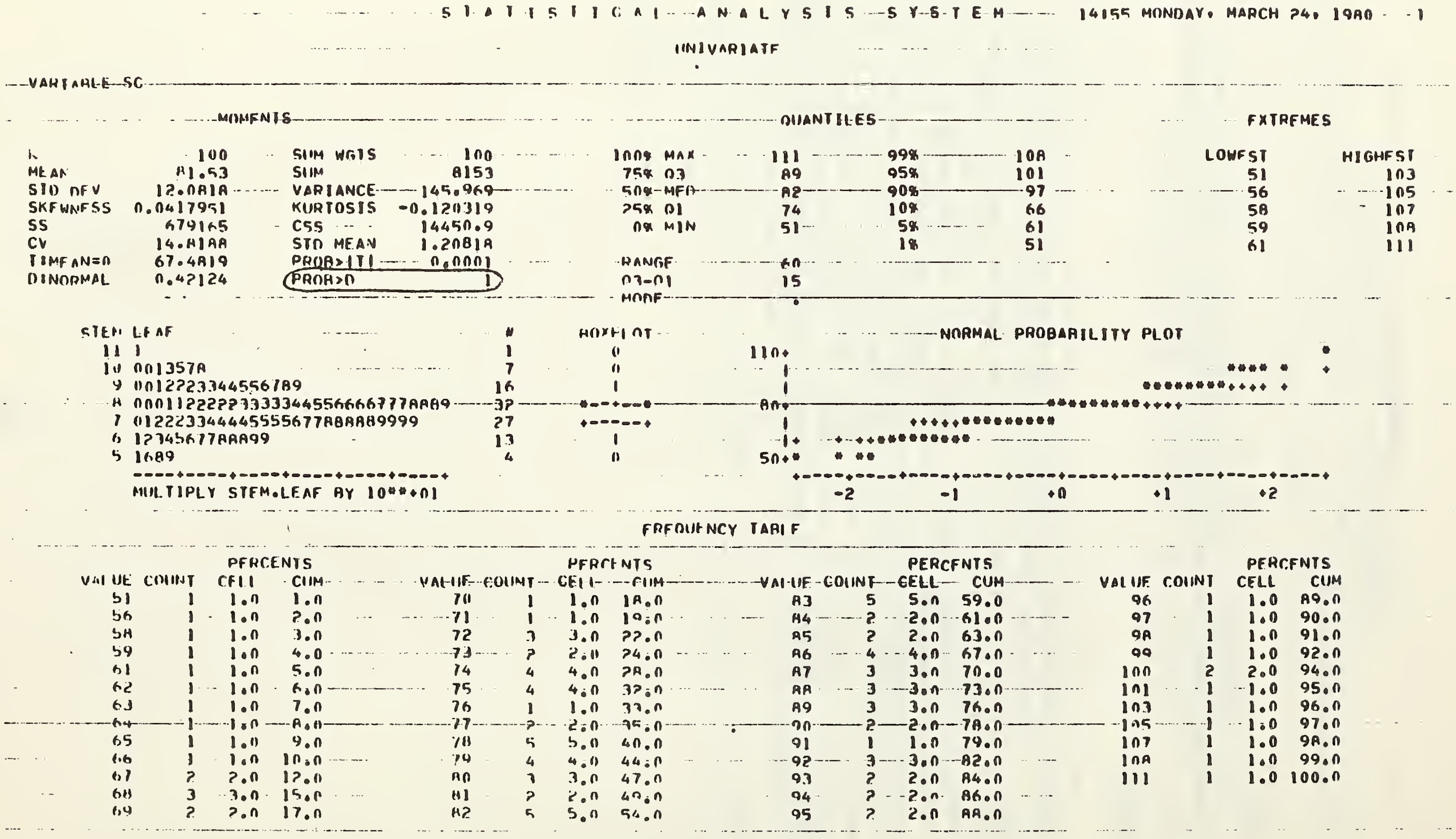
EXAMPLE 4a
Kolmogorov-Smirnov Test for Normality

Problem:

Use the UNIVARIATE procedure from SAS (1979), which applies the Kolmogorov-Smirnov test when $n > 50$, to test the specific conductance and fecal coliform populations presented in Table 3 for normality.

Solution:

The SAS output for this example is presented in Tables 8 and 9. The Kolmogorov-Smirnov test for normality results in a D-statistic (bottom of column 1 in the upper left portion of the output) and the probability of a larger D (bottom of column 2 in the upper left portion of the output). The PROB > D ranges from 0 to 1 which is the likelihood of obtaining a D value greater than the one printed. In other words, as long as PROB > D is equal to or greater than 0.95 we can readily accept the hypothesis that the observed distribution is no different from a normal distribution. In the case of the specific conductance data (Table 8) we would accept it as normally distributed since PROB > D = 1 while in the case of the fecal coliform data (Table 9) we would reject it as being normally distributed (PROB > D = 0.0256269). If we accept the fecal coliform data as normally distributed there is a 97.4% chance we have made the wrong decision.



```
                         S T A T I S T I C A L   A N A L Y S I S   S Y S T E M        14:55 MONDAY, MARCH 24, 1980   1

                                                   UNIVARIATE

VARIABLE=SC

              MOMENTS                                              QUANTILES                                      EXTREMES

N                100    SUM WGTS         100       100% MAX    111      99%      108            LOWEST        HIGHEST
MEAN           81.53    SUM             8153        75% Q3      89      95%      101               51            103
STD DEV      12.0818    VARIANCE     145.969        50% MED     82      90%       97               56            105
SKEWNESS   0.0417951    KURTOSIS   -0.120319        25% Q1      74      10%       66               58            107
SS            679165    CSS          14450.9         0% MIN     51       5%       61               59            108
CV           14.8188    STD MEAN     1.20818                             1%       51               61            111
T:MEAN=0     67.4819    PROB>|T|      0.0001       RANGE        60
D:NORMAL     0.42124    PROB>D             1       Q3-Q1        15
                                                   MODE

   STEM LEAF                             #    BOXPLOT                     NORMAL PROBABILITY PLOT
     11 1                                1       0          110+                                                        *
     10 0013578                          7       0             |                                                ****  *    +
      9 0012223344556789                16       |             |                                       ********++++  +
      8 00011222223333344556666777889   32    *--+--*        80+                         **********++++
      7 012223344445555677888889999     27    +-----+          |              +++++*********
      6 12345677888999                  13       |            -|+    +-+++*********
      5 1689                             4       0          50+*    * **
        ----+----+----+----+----+----+                        +----+----+----+----+----+----+----+----+----+----+
   MULTIPLY STEM.LEAF BY 10**+01                                  -2        -1        +0        +1        +2

                                                FREQUENCY TABLE

               PERCENTS                          PERCENTS                          PERCENTS                          PERCENTS
VALUE COUNT   CELL   CUM      VALUE COUNT   CELL   CUM      VALUE COUNT   CELL   CUM      VALUE COUNT   CELL   CUM
   51     1    1.0    1.0        70     1    1.0   18.0        83     5    5.0   59.0        96     1    1.0   89.0
   56     1    1.0    2.0        71     1    1.0   19.0        84     2    2.0   61.0        97     1    1.0   90.0
   58     1    1.0    3.0        72     3    3.0   22.0        85     2    2.0   63.0        98     1    1.0   91.0
   59     1    1.0    4.0        73     2    2.0   24.0        86     4    4.0   67.0        99     1    1.0   92.0
   61     1    1.0    5.0        74     4    4.0   28.0        87     3    3.0   70.0       100     2    2.0   94.0
   62     1    1.0    6.0        75     4    4.0   32.0        88     3    3.0   73.0       101     1    1.0   95.0
   63     1    1.0    7.0        76     1    1.0   33.0        89     3    3.0   76.0       103     1    1.0   96.0
   64     1    1.0    8.0        77     2    2.0   35.0        90     2    2.0   78.0       105     1    1.0   97.0
   65     1    1.0    9.0        78     5    5.0   40.0        91     1    1.0   79.0       107     1    1.0   98.0
   66     1    1.0   10.0        79     4    4.0   44.0        92     3    3.0   82.0       108     1    1.0   99.0
   67     2    2.0   12.0        80     3    3.0   47.0        93     2    2.0   84.0       111     1    1.0  100.0
   68     3    3.0   15.0        81     2    2.0   49.0        94     2    2.0   86.0
   69     2    2.0   17.0        82     5    5.0   54.0        95     2    2.0   88.0
```

Table 8. SAS output for the specific conductance portion of Example 4a.

```
                    S T A T I S T I C A L   A N A L Y S I S   S Y S T E M          14155 MONDAY, MARCH 24, 1980     2
                                                     UNIVARIATE

VARIABLE FC

                  MOMENTS                                              QUANTILES                                  EXTREMES

N                100     SUM WGTS         100          100% MAX     75        99%     66.9999              LOWEST     HIGHEST
MEAN           23.55     SUM             2355           75% Q3      27        95%          49                   7          54
STD DEV       12.546     VARIANCE     157.402           50% MED     20        90%          38                   8          59
SKEWNESS     1.87556     KURTOSIS      4.1614           25% Q1      15        10%          12                   9          61
SS             71043     CSS          15582.8            0% MIN      7         5%          10                  10          67
CV           53.2738     STD MEAN      1.2546                                  1%           7                  10          75
T:MEAN=0      18.771     PROB>|T|      0.0001          RANGE        68
DNORMAL      1.47602     PROB>D     0.0256269          Q3-Q1        12
                                                       MODE         16

  STEM LEAF                                     #    BOXPLOT                       NORMAL PROBABILITY PLOT
     7 5                                        1       *       70+                                                  *
     6 17                                       2       *         |                                            *  *
     5 49                                       2       *         |                                          **
     4 04569                                    5       *         |                                       ***         +
     3 001122378                                9       0         |                                  +****+++++ +
     2 00001111222233334444555666777888999     35    +--+--+      |                       ++++***********
     1 0011122233334444455555666666777778888899  43  *-----*      |              ****************
     0 789                                      3       |        0+*   * ***
       ----+----+----+----+----+----+----+----+                  +----+----+----+----+----+----+----+----+----+----+
  MULTIPLY STEM.LEAF BY 10**+01                                      -2        -1        +0        +1        +2

                                                     FREQUENCY TABLE

               PERCENTS                          PERCENTS                          PERCENTS                          PERCENTS
  VALUE COUNT  CELL   CUM       VALUE COUNT  CELL   CUM       VALUE COUNT  CELL   CUM       VALUE COUNT  CELL   CUM
      7     1   1.0   1.0          17     5   5.0  36.0          27     3   3.0  75.0          44     1   1.0  92.0
      8     1   1.0   2.0          18     5   5.0  41.0          28     3   3.0  78.0          45     1   1.0  93.0
      9     1   1.0   3.0          19     5   5.0  46.0          29     3   3.0  81.0          46     1   1.0  94.0
     10     2   2.0   5.0          20     4   4.0  50.0          30     2   2.0  83.0          49     1   1.0  95.0
     11     3   3.0   8.0          21     4   4.0  54.0          31     2   2.0  85.0          54     1   1.0  96.0
     12     3   3.0  11.0          22     4   4.0  58.0          32     2   2.0  87.0          59     1   1.0  97.0
     13     4   4.0  15.0          23     4   4.0  62.0          33     1   1.0  88.0          61     1   1.0  98.0
     14     5   5.0  20.0          24     4   4.0  66.0          37     1   1.0  89.0          67     1   1.0  99.0
     15     5   5.0  25.0          25     3   3.0  69.0          38     1   1.0  90.0          75     1   1.0 100.0
     16     6   6.0  31.0          26     3   3.0  72.0          40     1   1.0  91.0
```

Table 9. SAS output for the fecal coliform portion of Example 4a.

EXAMPLE 4b
Shapiro-Wilk Test for Normality

Problem:

Use the UNIVARIATE procedure from SAS (1979) which applies the Shapiro-Wilk test when $n \leq 50$, to test the specific conductance and fecal coliform samples used in Example 1 for normality.

Solution:

The SAS output for this example is presented in Tables 10 and 11. The procedure produces the Shapiro-Wilk W-statistic and the probability of a smaller W (located in the upper left of the output at the bottom of columns 1 and 2). The PROB < W ranges from 0 to 1 which is the likelihood of obtaining a W value less than the one calculated. As long as the PROB < W is 0.95 or greater we can readily accept the hypothesis that the sample has been obtained from a population which is normally distributed. In the case of the specific conductance sample, we accept it as coming from a normally distributed population since PROB < W = 0.95. The fecal coliform sample, on the other hand, has a PROB < W = 0.01 which results in the rejection of the hypothesis that it was obtained from a normally distributed population.

```
                              S T A T I S T I C A L   A N A L Y S I S   S Y S T E M          13:33 MONDAY, MARCH 24, 1980    1

                                                     UNIVARIATE

VARIABLE=SC

                 MOMENTS                                                     QUANTILES                                       EXTREMES

N                 15        SUM WGTS          15        100% MAX     101          99%      99.65              LOWEST      HIGHEST
MEAN            81.6        SUM             1224         75% Q3     84.5          95%      94.25                  65           84
STD DEV       8.6586        VARIANCE     74.9714         50% MED    80.5          90%         91                  73           86
SKEWNESS    0.409897        KURTOSIS     1.01194         25% Q1    76.25          10%         69                  74           90
SS            100928        CSS           1049.6          0% MIN      65           5%         65                  77           92
CV            10.611        STD MEAN     2.23564                                   1%         65                  77          101
T:MEAN=0     36.4996        PROB>|T|      0.0001        RANGE         36
W:NORMAL    0.976867        PROB<W          0.95        Q3-Q1       8.25
                                                        MODE           .

       STEM LEAF                      #       BOXPLOT                               NORMAL PROBABILITY PLOT
         10 1                         1          *                   100+                                            *
          9 02                        2          |                      |                                    *   +
          8 013346                    6       *--+--*                 80+                  * * ** * * *   +
          7 34778                     5       +-----+                   |      +    *  * *** *
          6 5                         1   °      0                    60+     *
            ----+----+----+----+                                         +----+----+----+----+----+----+----+----+----+----+
          MULTIPLY STEM.LEAF BY 10**+01                                          -2        -1        +0        +1        +2

                                                     FREQUENCY TABLE

               PERCENTS                          PERCENTS                          PERCENTS                          PERCENTS
  VALUE  COUNT  CELL   CUM       VALUE  COUNT  CELL   CUM       VALUE  COUNT  CELL   CUM       VALUE  COUNT  CELL   CUM
     65      1   6.7   6.7          78      1   6.7  40.0          84      1   6.7  73.3         101      1   6.7 100.0
     73      1   6.7  13.3          80      1   6.7  46.7          86      1   6.7  80.0
     74      1   6.7  20.0          81      1   6.7  53.3          90      1   6.7  86.7
     77      2  13.3  33.3          83      2  13.3  66.7          92      1   6.7  93.3
```

Table 10. SAS output for the specific conductance portion of Example 4b.

```
                    S T A T I S T I C A L   A N A L Y S I S   S Y S T E M        13:33 MONDAY, MARCH 24, 1980     2

                                                        UNIVARIATE

VARIABLE FC

               MOMENTS                                                    QUANTILES                                   EXTREMES

N                   15     SUM WGTS          15          100% MAX      49        99%         46             LOWEST      HIGHEST
MEAN           22.1333     SUM              332           75% Q3    23.25        95%         34                 12           23
STD DEV        8.79827     VARIANCE     77.4095           50% MED    19.5        90%       28.5                 15           24
SKEWNESS       2.16738     KURTOSIS     6.14815           25% Q1    16.75        10%       13.5                 16           28
SS                8432     CSS          1083.73            0% MIN      12         5%         12                 17           29
CV             39.7512     STD MEAN      2.2717                                   1%         12                 17           49
T:MEAN=0       9.74306     PROB>|T|      0.0001          RANGE         37
W:NORMAL      0.796151     PROB<W          0.01          Q3-Q1        6.5
                                                         MODE           .

      STEM LEAF                     #          BOXPLOT                         NORMAL PROBABILITY PLOT
         4 9                        1             *                40+                                                  *
         3                                                          |                                                  +
         2 0233489                  7          +--+--+              |                  ++  +  ****  *  *    *
         1 2567789                  7          *-----*            10+           *        *     *-***-**-*
           ----+----+----+----+                                      +----+----+----+----+----+----+----+----+----+----+
      MULTIPLY STEM.LEAF BY 10**+01                                       -2        -1        +0        +1        +2

                                                   FREQUENCY TABLE

                    PERCENTS                          PERCENTS                           PERCENTS                          PERCENTS
     VALUE  COUNT   CELL    CUM        VALUE  COUNT   CELL    CUM        VALUE  COUNT   CELL    CUM        VALUE  COUNT   CELL    CUM
        12      1    6.7    6.7           18      1    6.7   40.0           23      2   13.3   73.3           49      1    6.7  100.0
        15      1    6.7   13.3           19      1    6.7   46.7           24      1    6.7   80.0
        16      1    6.7   20.0           20      1    6.7   53.3           28      1    6.7   86.7
        17      2   13.3   33.3           22      1    6.7   60.0           29      1    6.7   93.3
```

Table 11. SAS output for the fecal coliform portion of Example 4b.

5.1.4 Transformations

Now that we know how to test our data for normality, the question arises, "What do we do if our data are not normally distributed?" Basically, there are three options available: (1) determine what theoretical frequency distribution it follows, such as the binomial or the Poisson, and proceed accordingly; or (2) use nonparametric statistical methods; or (3) transform the data.

We suggest you apply a data transformation first unless you have strong reason to believe the data follow another theoretical frequency distribution. In general, a log (X) transformation will normalize a water quality data set composed of continuous or measured data. This is especially true if the data are skewed to the right. In the case where zero values are present, then the transformation to use is log (X + 1). This avoids the problem of taking the log of zero. There are other transformations available; however, the log (X) or log (X + 1) will generally correct your problem of nonnormality. If it does not, then we suggest you consult a statistician before proceeding with your analysis. Example 5 illustrates the use of the log (X) transformation.

EXAMPLE 5
Log (X) Transformation

Problem:

Transform the fecal coliform data contained in Table 3 by log (X) and test it for normality, using SAS (1979) or SPSS (1975).

Solution:

1. Transform each FC_i in Table 3 to log (FC_i).

FC	Log (FC)	FC	Log (FC)	FC	Log (FC)	FC	Log (FC)	FC	Log (FC)
7	0.8451	15	1.1761	18	1.2553	23	1.3617	29	1.4624
8	0.9031	15	1.1761	19	1.2798	23	1.3617	30	1.4771
9	0.9542	15	1.1761	19	1.2798	24	1.3802	30	1.4771
10	1.0000	15	1.1761	19	1.2798	24	1.3802	31	1.4914
10	1.0000	15	1.1761	19	1.2798	24	1.3802	31	1.4914
11	1.0414	16	1.2041	19	1.2798	24	1.3802	32	1.5051
11	1.0414	16	1.2041	20	1.3010	25	1.3979	32	1.5051
11	1.0414	16	1.2041	20	1.3010	25	1.3979	33	1.5185
12	1.0792	16	1.2041	20	1.3010	25	1.3979	37	1.5682
12	1.0792	16	1.2041	20	1.3010	26	1.3979	38	1.5798
12	1.0792	16	1.2041	21	1.3222	26	1.4150	40	1.6021
13	1.1139	17	1.2304	21	1.3222	26	1.4150	44	1.6435
13	1.1139	17	1.2304	21	1.3222	27	1.4314	45	1.6532
13	1.1139	17	1.2304	21	1.3222	27	1.4314	46	1.6628
13	1.1139	17	1.2304	22	1.3424	27	1.4314	49	1.6902
14	1.1461	17	1.2304	22	1.3424	28	1.4472	54	1.7324
14	1.1461	18	1.2553	22	1.3424	28	1.4472	59	1.7709
14	1.1461	18	1.2553	22	1.3424	28	1.4472	61	1.7853
14	1.1461	18	1.2553	23	1.3617	29	1.4624	67	1.8261
14	1.1461	18	1.2553	23	1.3617	29	1.4624	75	1.8751

2. The UNIVARIATE procedure from SAS (1979) was used to test the transformed data for normality. If SPSS (1975) was selected, the NPAR TESTS subprogram would have been applied. The SAS output is presented in Table 12. Since the PROB > D = 1, we accept the hypothesis that the data come from a population following a normal distribution.

It should be noted that prior to the log (X) transformation, the fecal coliform population was not normally distributed (Example 4a). However, the transformation made it conform to the normal frequency distribution. This will generally occur when the data are moderately skewed, as in this example problem.

```
                              S T A T I S T I C A L   A N A L Y S I S   S Y S T E M    10:47 WEDNESDAY, MARCH 26, 1980    1
                                                     UNIVARIATE

VARIABLE LFC

              MOMENTS                                                  QUANTILES                                          EXTREMES

N                 100     SUM WGTS          100     100% MAX  1.87506       99%      1.82607             LOWEST       HIGHEST
MEAN          1.32325     SUM           132.325      75% Q3   1.43136       95%       1.6902           0.845098       1.73239
STD DEV      0.200608     VARIANCE    0.0402436      50% MED  1.30103       90%      1.57978            0.90309       1.77085
SKEWNESS     0.398642     KURTOSIS     0.333874      25% Q1   1.17609       10%      1.07918           0.954242       1.78533
SS            179.084     CSS           3.98412       0% MIN 0.845098        5%            1                  1       1.82607
CV            15.1602     STD MEAN    0.0200609                              1%     0.845098                  1       1.87506
T:MEAN=0       65.962     PROB>|T|       0.0001     RANGE     1.02996
D:NORMAL      0.54667     PROB>D              1     Q3-Q1    0.255272
                                                    MODE      1.20412

  STEM LEAF                                          #   BOXPLOT                          NORMAL PROBABILITY PLOT
    18 38                                            2      *          1.8+                                                  *   *
    16 04569379                                      8      0             |                                        *******     +
    14 000111333555666889911278                     24   +-----+          |                                 *********+++++++ +
    12 00000333336666688888000002222444446666888    41   *--+--*          |                     ++**************++
    10 0044488811115555588888                       22   +-----+          |+    + *************
     8 505                                           3      0          0.8+*    * *
       ----+----+----+----+----+----+----+----+                           +----+----+----+----+----+----+----+----+----+----+
                                                                                -2        -1        +0        +1        +2

                                                          FREQUENCY TABLE

               PERCENTS                         PERCENTS                           PERCENTS                           PERCENTS
   VALUE COUNT  CELL   CUM     VALUE COUNT  CELL   CUM     VALUE COUNT  CELL   CUM     VALUE COUNT  CELL   CUM
0.845098     1   1.0   1.0   1.23045     5   5.0  36.0   1.43136     3   3.0  75.0   1.64345     1   1.0  92.0
 0.90309     1   1.0   2.0   1.25527     5   5.0  41.0   1.44716     3   3.0  78.0   1.65321     1   1.0  93.0
0.954242     1   1.0   3.0   1.27875     5   5.0  46.0    1.4624     3   3.0  81.0   1.66276     1   1.0  94.0
       1     2   2.0   5.0   1.30103     4   4.0  50.0   1.47712     2   2.0  83.0    1.6902     1   1.0  95.0
 1.04139     3   3.0   8.0   1.32222     4   4.0  54.0   1.49136     2   2.0  85.0   1.73239     1   1.0  96.0
 1.07918     3   3.0  11.0   1.34242     4   4.0  58.0   1.50515     2   2.0  87.0   1.77085     1   1.0  97.0
 1.11394     4   4.0  15.0   1.36173     4   4.0  62.0   1.51851     1   1.0  88.0   1.78533     1   1.0  98.0
 1.14613     5   5.0  20.0   1.38021     4   4.0  66.0    1.5682     1   1.0  89.0   1.82607     1   1.0  99.0
 1.17609     5   5.0  25.0   1.39794     3   3.0  69.0   1.57978     1   1.0  90.0   1.87506     1   1.0 100.0
 1.20412     6   6.0  31.0   1.41497     3   3.0  72.0   1.60206     1   1.0  91.0
```

Table 12. SAS output for Example 5.

5.2 Student's t-Distribution

The student's t-distribution was introduced by William Gosset who wrote several statistical papers under the pseudonym of "Student." This distribution was developed for problems involving the sample mean when the population variance is not known and the sample size is small ($n < 30$). With this distribution, we can make statistical statements about the population mean using only the sample variance and the sample mean.

If our sample has been randomly drawn from a normal population with mean μ and variance σ, then the statistic

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \tag{6}$$

is distributed as the student's t-distribution with "n-1" degrees of freedom. $s_{\bar{X}}$ denotes the standard error of the mean. A reliable estimate of the standard error of the mean is obtained by:

$$s_{\bar{X}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2/n-1}{n}}$$

The theoretical t-distribution is similar to the normal distribution in that it extends from negative to positive infinitely but differs in shape depending on the number of degrees of freedom (ν). As the number of observations approaches infinity, the student's t-distribution approaches the normal distribution with mean 0 and variance 1.

Table A-2 lists the percent of the area in both tails of the curve using degrees of freedom (ν) and probability (α) as arguments. It is important that you remember this is a two-tailed table. Consequently, a probability of 0.05 means that 0.025 of the area is in each tail. For a single-tail test, you need to halve the probability argument prior to entering Table A-2. Values of t are commonly denoted as t_{α} (ν) = area;

for example, the t-value of $\alpha = 0.05$ and $\nu = 8$ is $t_{.05(8)} = 2.306$. It is very important that you understand how to use Table A-2 since you will use it often.

5.3 The Chi-square Distribution

The chi-square distribution has application for problems involving the sample variance. If a sample is randomly drawn from a normal population with mean μ and variance σ, then the statistic

$$\chi^2 = \frac{\Sigma(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \qquad (7)$$

follows the chi-squared distribution with "n-1" degrees of freedom. It should be apparent from Equation 7 that there is a relation between χ^2 and s^2, which is similar to the relation between t and X just discussed.

The chi-square distribution is similar to the t-distribution in that for each number of degrees of freedom, there is a specific chi-square curve. However, unlike the normal and t-distributions, the chi-square cannot be negative since it involves the square of the sum of squares.

In general, only a few values for each of the many chi-square curves are tabulated. Table A-3 presents chi-square values for given probability levels and degrees of freedom. The table is designed so that we can determine the likelihood that a calculated χ_s^2 will exceed the theoretical χ_α^2 in other words

$$P(\chi_s^2 > \chi_\alpha^2) = \alpha$$

5.4 The F Distribution

The chi-square distribution allows us to test various statements concerning a single variance. However, if we want to test statements concerning the variances of two populations, we must use the F statistic.

It is given by the formula

$$F = \frac{s_1^2}{s_2^2} \tag{8}$$

and has a sampling distribution called the F distribution. There are two sample variances involved, s_1^2 and s_2^2 representing the variances of populations 1 and 2, respectively. Associated with each variance is its degrees of freedom, ν_1 (n_1-1) for the numerator and ν_2 (n_2-1) for the denominator. Appendix Table A-4 gives critical values of F. The table is entered with three values, α, ν_1, and ν_2. Critical F values are commonly denoted by $F_{\alpha(\nu_1,\nu_2)}$. The percentages listed in the table refer to the proportion of the area under the curves to the right of the values given in the tables. For example, for $F_{.01(10,12)} = 4.30$, one percent of the area is to the right of 4.30.

6.0 Confidence Limits About the Sample Mean

When we estimate a population parameter with a sample statistic, such as μ with $\bar{X}$, we should ask: "How reliable is our estimate?" Since the real values of the population parameters in our various water quality studies will always remain unknown to us, we cannot evaluate our estimate with a direct comparison. However, with the statistical methods we now have available to us, we can predict the reliability of our estimate by setting a confidence interval about it.

A confidence interval consists of two confidence limits which set an upper (U) and lower (L) bound to the interval. The level of confidence is expressed as a probability (P) or percent and indicates the likelihood that the interval obtained from a particular sample brackets the true parameter value (P'). This concept can be expressed as

$$P\ (L \leq P' \leq U) = 1 - \alpha$$

where α denotes the level of reliability. In other words, there is a $1-\alpha$ chance that the sample is representative of the population and that the confidence interval covers the population parameter.

Confidence limits for the sample statistics $\bar{X}$ and s^2 are determined using the t and χ^2 distributions. Confidence limits for a mean of a normally distributed population whose standard deviation is not known can be determined from a random sample using the equations:

$$L = \bar{X} - t_{\alpha(n-1)} s_{\bar{X}} \tag{9}$$

and

$$U = \bar{X} + t_{\alpha(n-1)} s_{\bar{X}} \tag{10}$$

An application of setting confidence limits is illustrated in Example 6.

EXAMPLE 6
Setting Confidence Limits for the Population Mean

Problem:

Twenty-one water samples were collected in a random manner at the mouth of Gypsum Creek and analyzed for total dissolved solids (TDS). The data were tested and found to be normally distributed with a mean TDS of 517 mg/l and a standard deviation of 67 mg/l. Determine the 95% and 99% confidence limits for the population mean, μ .

Solution:

Case I. Determination of the 95% confidence interval about the mean.

a. Lower Limit

$$L = \bar{X} - t_{.05(n-1)} \frac{s}{\sqrt{n}}$$

$$L = \bar{X} - t_{.05(20)} \frac{s}{\sqrt{n}}$$

$$L = 517 - 2.086 \left(\frac{67}{\sqrt{21}}\right) = 517 - 30.5$$

$$L = 486.5 \text{ mg/l}$$

b. Upper Limit

$$U = \bar{X} + t_{.05(n-1)} \frac{s}{\sqrt{n}}$$

$$U = 517 + 30.5$$

$$U = 547.5 \text{ mg/l}$$

Conclusion: There is a 95% chance that the population mean, μ, will be covered by the interval 486.5 to 547.5 mg/l.

Case II: Determine the 99% confidence interval about the mean.

a. Lower Limit

$$L = \bar{X} - t_{.01(n-1)} \frac{s}{\sqrt{n}}$$

$$L = 517 - 2.845 \left(\frac{67}{\sqrt{21}}\right) = 517 - 41.6$$

$$L = 475.4 \text{ mg/l}$$

b. Upper Limit

$$U = \bar{X} + t_{.01(n-1)} \frac{s}{\sqrt{n}}$$

$$U = 517 + 41.6$$

$$U = 558.6 \text{ mg/l}$$

Conclusion: There is a 99% chance that the interval 475.4 to 558.6 mg/l will cover the population mean. Note that the 99% confidence interval is larger than the 95% one.

7.0 Hypothesis Testing

Hypothesis testing plays an integral role in statistical decision-making. A statistical hypothesis is simply a statement made about the expected results of a study, such as $\mu = \mu_0$, or $\mu_1 < \mu_2$. The fundamental hypothesis is the null hypothesis, denoted by H_0. In some cases, the null hypothesis is one of no difference. In others, it states that a given parameter does not exceed some standard or other value. In practice, we may seriously doubt the truth of H_0 from the moment it is proposed. However, its purpose is to give us a starting point from which we can calculate a meaningful test statistic and come to an objective decision.

Any hypothesis that differs from the null hypothesis is called an alternate hypothesis, denoted by H_a.

Once we have established our hypotheses, H_0 and H_a, we proceed to sample the population in question. These data are then analyzed statistically and a decision is made to either accept or reject the H_0. Unfortunately, there is no guarantee that our decision will be correct. In fact, there are two mistakes possible: (1) if the stated hypothesis is true, we might erroneously call it false (Type I error) and (2) if the stated hypothesis is false, we might erroneously call it true (Type II error). Our decision to accept or reject a hypothesis will be based on two factors: (1) the information we obtain from our sample and (2) the risk that we are willing to take that our decision may be wrong.

The Type I error is generally expressed as a probability and is denoted by "α". When it is expressed as a percentage, it is termed significance level. Consequently, a Type I error of $\alpha = 0.05$ is equivalent to a significance level of 5%. The level of significance should be established prior to data collection with the following guidelines in mind.

If it is a matter of serious concern when a true hypothesis is rejected, such as evidence to shut down a logging operator, the risk of making this error, α, should be small. However, if it is important that the hypothesis be rejected when there is slight evidence against it, such as contamination of a public water supply, we should choose a larger α.

The concept of rejection or acceptance of the null hypothesis can best be illustrated with an example. Suppose we hypothesize that a body of water has a mean TDS concentration equal to or less than 80 mg/l. Therefore, H_0: $\mu \leq 80$ mg/l and H_a: $\mu > 80$ mg/l are the null and alternate hypotheses. We choose a significance level of 5% as the basis for rejection of H_0. It is assumed that the TDS in the water body are normally distributed. A sample of 14 observations is collected from the water body in a random manner and found to have a mean of 89 mg/l and a variance of 36. The question now is do we accept or reject H_0? Before we can do this we need to determine the critical region. As we will see later, the critical point in this case is defined by $\bar{X} + t_{.05(13)s}$ which is equal to 83.5 ($80 + 2.16\sqrt{36/14}$). The critical region is illustrated in Figure 7. Since 89 falls to the right of 82.6, we reject H_0 which means it is not very likely that the sample obtained came from a water body with a mean TDS of 80 mg/l.

When H_0 has been rejected at a given significance level, we say the sample is <u>significantly different</u>. The degree of statistical significance is a function of the probability level at which the difference is detected (Table 13).

The significant levels of 5, 1 and 0.1% are arbitrary. However, in the case of small samples ($n < 30$), it is not likely that H_0 will be rejected if these levels are required, unless a very large difference

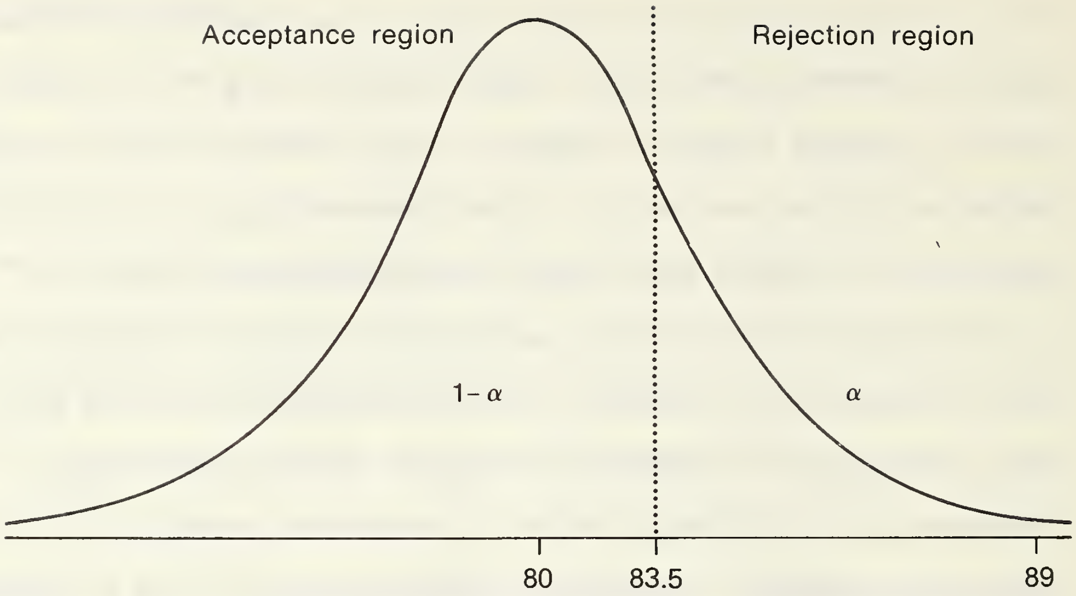


Figure 7. Rejection and acceptance regions.

exists. If this is the case, a significance level of 10% may be more appropriate (Steel and Torrie, 1960). If the null hypothesis is not rejected, it is considered <u>not significant</u> and is denoted by "ns".

TABLE 13. Probability levels for various degrees of significant differences.

Probability Level	Degree of Significance	Superscript Notation 1/
$.01 < p < .05$	Significant	*
$.001 < p < .01$	Highly Significant	**
$p < 0.001$	Very Significant	***

1/ Superscript notation is commonly used with a test statistic to denote the degree of significance, such as t = 2.05** to denote a highly significant difference.

At this point, we would like to caution you about your use of the word "significant". Sokal and Rohlf (1969) summarize this point very well.

> "Since statistical significance has special technical meaning, H_0 rejected at $P \leq \alpha$, we should use the adjective significant only in this sense; its use in scientific papers and reports, unless such a technical meaning is clearly implied, should be discouraged. For general descriptive purposes, synonyms such as important, meaningful, marked, noticeable, and others can serve to underscore differences and effects".

The Type II error, which is accepting the null hypothesis when it is false, is commonly denoted by β. The probability of a β error is determined by the choice of α and distance between μ_1, and μ_2, as illustrated in Figure 8. It is apparent that in the case of a fixed sample size, a reduction of α will be accompanied by an increase in β. Consequently, we need to consider very carefully the consequences of the different types of error when we choose a level of significance (Table 14).

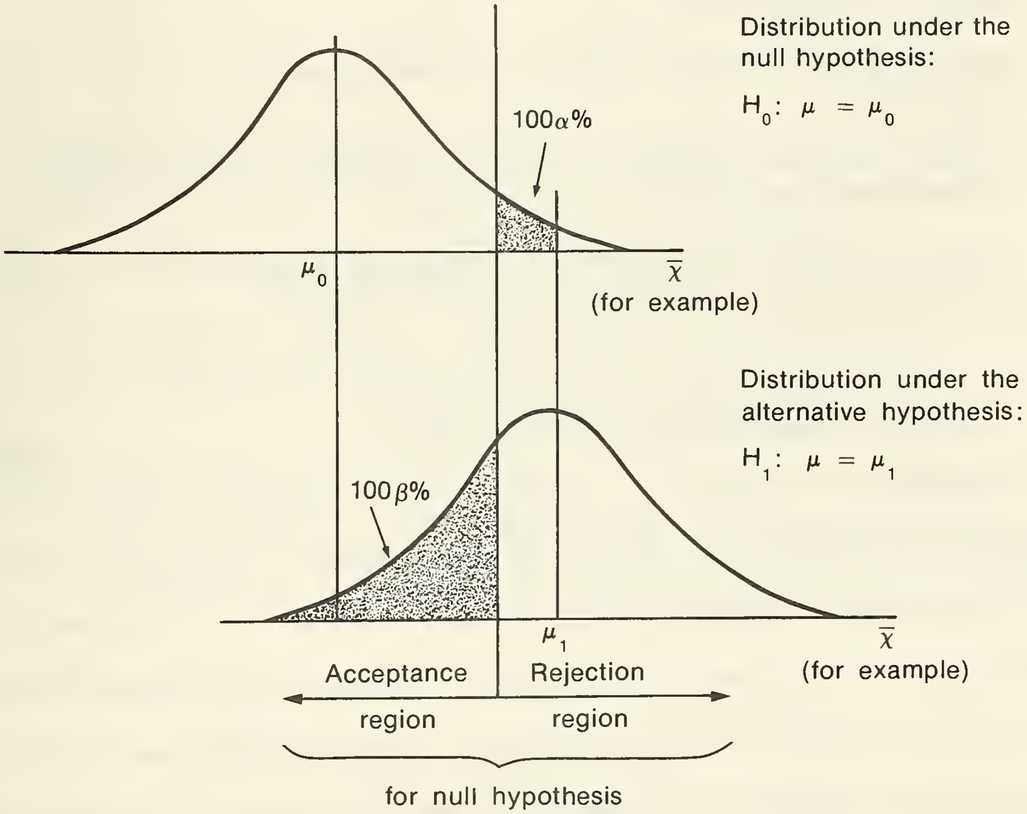


Figure 8. The relationship between the Type I and Type II errors (after Steel and Torrie, 1960).

Table 14. Statistical decisions and their outcomes, with special reference to the type of error (after Steel and Torrie, 1960).

Data is from a population for which		Decision relative to		Decision is	Probability
Null Hypothesis	Alternative hypothesis	H_o is	H_a is		should be
True	False	Accept	Reject	Right	High
True	False	Reject	Accept	Wrong; Type I error	Low
False	True	Accept	Reject	Wrong; Type II error	Low
False	True	Reject	Accept	Right	High

The basic steps for statistical hypothesis testing can be summarized as follows:

1. Establish H_o and H_a.

2. Ideally, specify α and β. However, in practice α and β are generally specified.

3. Determine the critical region for rejection of the null hypothesis.

4. Compute the test statistic from the observed values obtained through sampling.

5. Accept or reject the hypothesis depending on the position of the test statistic relative to the critical region.

8.0 Testing for Homogeneity of the Variance

Homogeneity, or equality, of variances in a group of samples is an important prerequisite for many of the statistical tests which follow. There are two basic procedures for testing equality of the variances: (1) when only two samples are involved we use the F test and (2) when there are more than two groups involved, Bartlett's test provides a means for evaluating the assumption. Examples 7a and 7b illustrate the computations involved in each of these procedures.

EXAMPLE 7a
Testing for Homogeneity Between Two Variances

Problem:

For the data given below determine whether the two variances can be considered equal at a significance level of 5%.

Given:

Sampling Station	Constituent	n	X	s^2
A	TDS	12	68	8
B	TDS	11	54	34

Solution:

1. Establish the hypothesis to be tested.

H_o: $\sigma_A^2 = \sigma_B^2$

H_a: $\sigma_A^2 \neq \sigma_B^2$

2. Select the significance level.

As stated in the problem, $\alpha = 0.05$

3. Use the test statistic $F_s = \frac{s_B^2}{s_A^2}$ to test the hypothesis.

$$F_s = \frac{s_B^2}{s_A^2} = \frac{34}{8} = 4.25$$

It should be noted here that since only the right tail of the F-distribution is tabulated in Table A-4, we calculate F_S as the ratio of the greater variance over the lesser one.

4. Define the critical region.

Since the test is two-tailed, we look up the critical value F where $\frac{\alpha}{2}$ is the Type I error and $\nu_B = n_B - 1$ and $\nu_A = n_A - 1$ are the degrees of freedom of samples B and A, respectively. From Table A-4 we find

$F_{0.025(10,11)} = 3.53$

5. Reject or accept the null hypothesis.

Since $F_S > F$ we reject the null hypothesis that the variances are equal at the 5% level of significance.

EXAMPLE 7b
Bartlett's Test of Homogeneity of Variances

Problem:

For the data given below determine whether the variances can be considered equal at a significance level of 5%.

Given:

Sampling station	n	s^2	Corrected sum of of squares (SS)
A	8	54.1	378.7
B	15	65.3	914.2
C	9	78.9	631.2
D	12	69.7	766.7

Solution:

1. Establish the hypothesis to be tested.

$$H_o: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$$

$$H_a: \sigma_A^2 \neq \sigma_B^2 \neq \sigma_C^2 \neq \sigma_D^2$$

2. Select the significance level.

From the problem statement, $\alpha = 0.05$.

3. Transform the variances to log (s^2) and compute $\Sigma\big((n_i - 1)\log(s_i^2)\big)$

Station	log (s^2)	$(n-1)$ log (s^2)
A	1.7332	12.132
B	1.8149	25.409
C	1.8971	15.177
D	1.8432	20.275
		$\Sigma =$ 72.993

4. Sum the degrees of freedom, i.e. $\Sigma(n_i - 1)$, and the corrected sum of squares.

$$\Sigma(n_i - 1) = 40$$

$$\Sigma(SS_i) = 2690.80$$

5. Determine the log of the pooled within-group variance, log (s^2).

$$\log(\bar{s}^2) = \log\left[\frac{\Sigma SS_i}{\Sigma(n_i - 1)}\right] = \log\left[\frac{2690.8}{40}\right] = 1.8278$$

6. Determine $\chi^2_{(a-1)}$ d.f. where a is the sum of the number of stations or groups considered; in this example a = 4.

$$\chi_s^2 = 2.3026 \left[\left((\log(\bar{s}^2)\Sigma(n_i - 1)\right) - \Sigma\left((n_i - 1)\log(\bar{s}_i^2)\right)\right]$$

$$\chi_s^2 = 2.3026 \left[1.8278(40) - 72.993\right] = 0.27598$$

Note: 2.3026 transforms the common logarithms to natural logarithms.

$$\chi_s^2 = 0.27598$$

7. Define the critical region and accept or reject H_0.

The critical chi-square is defined as $\chi^2_{\alpha(a-1)}$, which in this example (from Table Λ-3) is

$$\chi^2_{.05(3)} = 7.815$$

Since $\chi_s^2 < \chi^2_{.05(3)}$ we do not reject H_0.

Conclusion:

The data from stations A, B, C and D cannot be considered to have equal variances at the 5% level of significance.

It should be noted that the equation for χ_s^2 is biased slightly upward. If χ_s^2 is nonsignificant, the bias is not important. However, if the computed χ_s^2 is just a little above the threshold value for significance, a correction (C) for the bias should be applied as follows:

$$\chi_{s'}^2 = \frac{\chi_s^2}{C}$$

where $\chi_{s'}^2$ is the corrected χ_s^2 for the bias and

$$C = \frac{3(a-1) + \left[\Sigma\left(\frac{1}{n_i - 1}\right) - \frac{1}{\Sigma(n_i - 1)}\right]}{3(a-1)}$$

The SAS(BMDP) examples described by Ingwersen (1981a) show how to obtain the Bartlett chi-square test statistic of equal variances using SAS.

9.0 Comparison of Two Population Means or a Population Mean and a Constant

In many instances, we need to determine if a population mean is significantly different from an established value, such as a fixed water quality standard, or whether two populations are significantly different in an above-versus-below or treatment-versus-control study. In this section, three distinct cases are reviewed: (1) comparison of the population mean and a constant when the variance is unknown, (2) comparison of the means from two populations when the variance is unknown and the data are unpaired and (3) comparison of the means from two populations when the variance is unknown and the data are paired. In each case, a brief description of the utility of the technique, the procedure used in testing the hypothesis and a water quality example are presented.

Case I. Comparison of the population mean and a constant when the variance is unknown.

This particular case should be applied in situations where we wish to compare the mean value of a water quality constituent at a given sampling site with a specified constant, such as a nonvariable water quality standard. It is assumed that the data were collected in a random manner from a normally distributed population. The procedure for testing the hypothesis is as follows:

1. Establish H_0 and H_a.

   If we want to test that the population mean (μ) is equal to a specified constant (c), then H_0 and H_a are:

   $$H_o\colon \mu = c$$

   $$H_a\colon \mu \neq c.$$

   However, if we want to test that the population mean is not greater than the specified constant, then H_0 and H_a are:

   $$H_o\colon \mu \leq c$$

   $$H_a\colon \mu > c.$$

2. Test the data for normality.

3. Select the significance level (α).

4. Compute the test statistic. $t_s = \frac{\bar{X} - c}{s_{\bar{X}}}$.

5. Define the critical t value, t_C. In the instance where we are testing H_0: $\mu = c$, the test is two-tailed while for the situation where H_0: $\mu \leq c$ the test is one tailed.

6. Reject or accept the null hypothesis.

An illustration of Case I is presented in Example 8a.

EXAMPLE 8a
Comparison of the Population Mean and a Constant When the Variance is Unknown

---

Problem:

Twenty suspended solids samples were collected at a sampling station over three months. The concentrations of SS were determined and are as follows: 16, 29, 75, 7, 15, 24, 26, 12, 10, 23, 49, 22, 13, 14, 18, 17, 19, 31, 27 and 14. Is there reason to believe, at the 5% level of significance, that the mean SS concentration is significantly different from 28 mg/l?

Solution:

In general, SS data is strongly skewed. To test the assumption of normality, the data was subjected to the Shapiro-Wilk test using the UNIVARIATE procedure from SAS (1979). The result was a PROB < W = 0.01. Consequently, the data was transformed using Log (X) and retested for normality. The result of the log (X) transformation was a PROB < W = 0.90. This was much better, but there is still a 10% chance of Type I error if we accept the sample as being from a normal population. However, since the sample is small, we decide to accept it coming from a normally distributed population and proceed with the problem at hand.

1. Establish H_o and H_a.

$$H_o: \mu = \log(28)$$

$$H_a: \mu \neq \log(28)$$

2. Select the significance level.

From the problem statement, $\alpha = 0.05$.

3. Compute the test statistic, t_s.

$$t_s = \frac{\bar{X} - c}{s_{\bar{x}}}$$

$$t_s = \frac{1.2962 - 1.4472}{0.2348/\sqrt{20}} = -2.8760$$

4. Define the critical region.

Using Table A-2 we find:

$t < t_{.025(19)}$; therefore $t < 2.093$

and

$t > t_{.975(19)}$; therefore $t > 2.093$

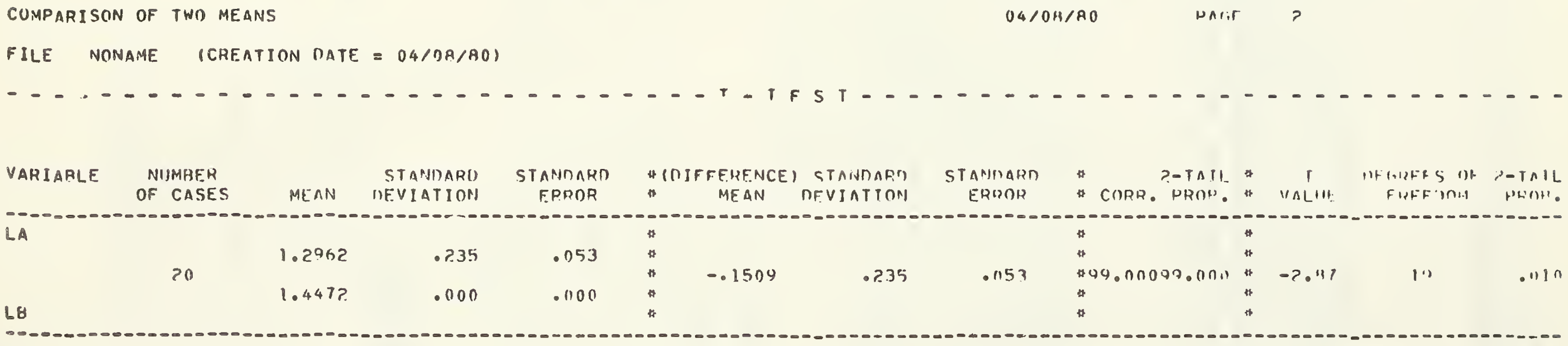
5. Reject or accept H_O.

Since t_s lies to the left of -2.093 we do not accept the null hypothesis.

6. Conclusion.

At the 5% level of significance, the mean SS concentration is significantly different from 28 mg/l.

This example problem can also be solved using either the MEANS (T,PRT) procedure from SAS (1979) or the T-TEST (PAIRS=) subprogram from SPSS (1975). The SPSS output for this problem is presented in Table 15. It should be noted that the mean values are in terms of log (X).



```
COMPARISON OF TWO MEANS                                                                              04/08/80          PAGE     2

FILE    NONAME    (CREATION DATE = 04/08/80)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T - T E S T - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

VARIABLE     NUMBER                STANDARD    STANDARD   *(DIFFERENCE) STANDARD    STANDARD   *       2-TAIL  *    T     DEGREES OF  2-TAIL
            OF CASES      MEAN    DEVIATION     ERROR     *      MEAN    DEVIATION    ERROR     * CORR. PROB.  *  VALUE    FREEDOM    PROB.
-----------------------------------------------------------------------------------------------------------------------------------------------
LA                                                        *                                     *              *
                        1.2962       .235        .053     *                                     *              *
                20                                        *    -.1509        .235       .053    *99.00099.000  *  -2.87       19        .010
                        1.4472       .000        .000     *                                     *              *
LB                                                        *                                     *              *
-----------------------------------------------------------------------------------------------------------------------------------------------
```

Table 15. SPSS output for Example 8a.

Case II: Comparison of the means from two populations when the variance is unknown and the data are unpaired.

In some water quality studies we want to know if two population means are statistically different. An example might be assessing the impact of a road when sampling above and below was performed or evaluating the effect of a treated watershed against a control watershed. In addition to the assumptions for Case I, it is assumed that the variances of both populations are equal, although unknown, and that the data are not paired. The procedure for testing the hypothesis is as follows:

1. Establish H_O and H_a.

$$H_o: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

2. Test the data for normalilty and homogeneity among variances.

3. Select the significance level (α).

4. Compute the test statistic, t_S, to test the hypothesis where

$$t_s = \frac{\bar{d}}{s_{\bar{d}}}, \quad \bar{d} = \bar{X}_1 - \bar{X}_2, \quad s_{\bar{d}} = \sqrt{s_w^2 \left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

and s_W which is the weighted variance, is determined by

$$s_w^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

5. Determine the critical t value, t_C. This is a two-tailed test.

$$t_c = \pm t_{1/2\alpha(n_1 + n_2 - 2)}$$

6. Reject or accept the hypothesis.

An example of Case II is given in Example 8b.

EXAMPLE 8b
Comparison of the Means from Two Populations When the Population Variance is Unknown and the Data Unpaired

---

Problem:

Water temperature (°F) was measured at the mouths of two different tributaries, A and B, to Whitefish Creek during the summer of 1980. The data obtained are summarized below.

Station A: 66, 59, 74, 60, 62, 69, 78, 71, 52, 78, 44, 50, 64

Station B: 61, 69, 67, 63, 39, 80, 63, 78, 47, 67, 72, 80, 41, 52, 64, 66,
74, 65, 67, 62

Is there a significant difference, $\alpha = 0.05$, between the population means of Stations A and B? It can be assumed that the populations are normally distributed and have equal variances.

Solution:

1. Establish H_0 and H_a.

$$H_o\colon \mu_1 = \mu_2$$

$$H_a\colon \mu_1 \neq \mu_2$$

2. Select the significance level.

From the problem statement we know $\alpha = 0.05$.

3. Compute the test statistic, t_s.

First, compute $\bar{X}$ and s for Stations A and B.

$$\bar{X}_A = 63.62$$
$$s_A = 10.62$$
$$\bar{X}_B = 63.85$$
$$s_B = 11.54$$

Now, determine t_s as follows:

$$t_s = \frac{\bar{d}}{s_{\bar{d}}}$$

$$\bar{d} = \bar{X}_A - \bar{X}_B$$

$$\bar{d} = 63.62 - 63.85 = -0.23$$

$$s_{\bar{d}} = \sqrt{s_w^2\left(\frac{n_A + n_B}{n_A n_B}\right)}$$

$$s_w^2 = \frac{s_A^2(n_A - 1) + s_B^2(n_B - 1)}{n_A + n_B - 2}$$

$$s_w^2 = \frac{(10.62)^2(12) + (11.54)^2(19)}{31}$$

$$s_w^2 = 125.3$$

$$s_{\bar{d}} = \sqrt{125.3\left(\frac{13 + 20}{13 \times 20}\right)}$$

$$t_s = \frac{-0.23}{3.988}$$

$$t_s = -0.058$$

4. Define the critical regions.

The critical regions for this example are defined by:

$t_c < t_{.025(31)}$

and

$t_c > t_{(1 - 0.025)(31)}$

Therefore, the rejection regions are:

$t < -2.042$

and

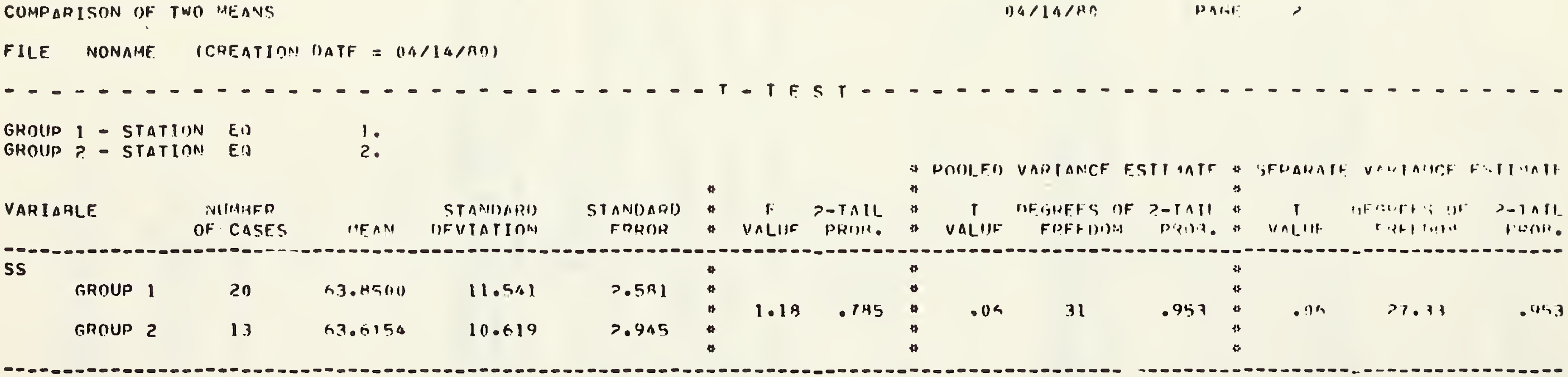
$t > 2.042$

5. Reject or accept the null hypothesis.

Since t_s lies between the t values of -2.042 and 2.042 we do not reject the null hypothesis.

Conclusion: There is no difference between the mean water temperatures at the mouths of streams A and B at the 5% level of significance.

This example problem could also be solved using the T-TEST program from SAS (1979) or the T-TEST (GROUPS=) subprogram from SPSS (1975). The SPSS output for the solution of example problem is presented in Table 16.



```
COMPARISON OF TWO MEANS                                                                                04/14/80        PAGE    2

FILE   NONAME   (CREATION DATE = 04/14/80)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - T - T E S T - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

GROUP 1 - STATION  EQ         1.
GROUP 2 - STATION  EQ         2.
                                                                                   * POOLED VARIANCE ESTIMATE * SEPARATE VARIANCE ESTIMATE
                                                                  *                *                          *
VARIABLE             NUMBER              STANDARD      STANDARD   *    F    2-TAIL *    T    DEGREES OF 2-TAIL *    T     DEGREES OF   2-TAIL
                    OF CASES      MEAN   DEVIATION        ERROR   *  VALUE   PROB. *  VALUE   FREEDOM    PROB. *  VALUE    FREEDOM      PROB.
---------------------------------------------------------------------------------------------------------------------------------------------
SS                                                                *                *                          *
      GROUP 1          20      63.8500     11.541         2.581   *                *                          *
                                                                  *   1.18    .785 *   .06       31       .953 *   .06      27.33        .953
      GROUP 2          13      63.6154     10.619         2.945   *                *                          *
                                                                  *                *                          *
---------------------------------------------------------------------------------------------------------------------------------------------
```

Table 16. SPSS output for Example 8b.

Case III: Comparisons of the means from two populations when the variance is unknown and the data are paired.

In some instances, extraneous factors which have no direct relation to the effect we are attempting to measure can cause significant difference between the means. We can overcome this problem somewhat by designing our sampling program so that observations are collected in pairs, i.e. when a sample is collected at Station A, one is also collected at Station B. In this type of a study, we are striving to obtain a pair of observations that are alike in all respects except the one we are trying to measure. This method has the same underlying assumptions as Case II, with the exception of equal variances for each population. In other words, with this method we need not assume the two population variances are equal. The procedure for testing the hypothesis is as follows:

1. Establish H_0 and H_a.

$$H_o: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

2. Test the data for normality.

3. Select the significance level (α).

4. Use the statistic, t_s, to test the hypothesis:

$$t_s = \frac{\bar{d}}{s_{\bar{d}}/\sqrt{n}}$$

where

$$\bar{d} = \bar{X}_1 - \bar{X}_2$$

and $s_{\bar{d}}$ is the standard deviation of the differences.

5. Determine the critical t value, t_c. This is a two tailed test.

$$t_c = \pm t_{(\alpha/2)(n-1)}$$

6. Compute the value of the test statistic, t_s, and accept or reject the hypothesis.

An example of Case III is presented in Example 8c.

EXAMPLE 8c
Comparison of the Means from Two Populations When the Population Variance is Unknown and the Data Paired

---

Problem:

Water samples were collected in pairs above (A) and below (B) a clearcut in western Oregon. The samples were analyzed for dissolved oxygen concentration in the field. The results are tabulated below. Determine if there is a significant difference ($\alpha = 0.05$) in the mean dissolved oxygen concentration above and below the clearcut. It can be assumed that both populations are normally distributed.

Dissolved Oxygen Concentration (mg/l)		Difference
Station A	Station B	d = A - B
6.2	5.2	1.0
6.5	5.4	1.1
6.8	5.3	1.5
7.0	5.7	1.3
6.9	5.6	1.3
7.0	6.2	0.8
6.8	5.7	1.1
6.7	5.6	1.1
6.8	5.8	1.0
6.2	5.6	0.6
		d = 10.08
		d = 1.08
		s_d = 0.257

Solution:

1. Establish H_0 and H_a.

$$H_o: \mu_A = \mu_B$$

$$H_a: \mu_A \neq \mu_B$$

2. Select the significance levels.

From the problem statement we know $\alpha = 0.05$.

3. Compute the test statistic, t_s.

$$t_s = \frac{\bar{d}}{s_{\bar{d}}/\sqrt{n}}$$

$$t_s = \frac{1.008}{0.257/\sqrt{10}}$$

$$t_s = 12.40$$

4. Define the critical regions

$t_c < t_{.025(9)}$

and

$t_c > t_{.975(9)}$

Therefore, the rejection regions are:

$t < -2.262$

and

$t > 2.262$

5. Reject or accept the null hypothesis.

Since t_s is greater than 2.262, the null hypothesis is rejected.

Conclusion: The dissolved oxygen concentration is significantly different above and below the clearcut at the 5% level.

This example problem can also be solved using the MEANS(T,PRT) procedure from SAS (1979) or the T-TEST (PAIRS=) subprogram from SPSS (1975). The SPSS output for the solution of this example is presented in Table 17.

```
                                                                                03/31/80       PAGE    2
COMPARISON OF TWO MEANS

FILE   NONAME   (CREATION DATE = 03/31/80)
- - - - - - - - - - - - - - - - - - - - - - - T - T E S T - - - - - - - - - - - - - - - - - - - - - - - -

VARIABLE    NUMBER                STANDARD   STANDARD  *(DIFFERENCE) STANDARD   STANDARD  *        2-TAIL *    T     DEGREES OF 2-TAIL
           OF CASES      MEAN    DEVIATION    ERROR    *     MEAN    DEVIATION    ERROR    * CORR.  PROB.  *  VALUE     FREEDOM   PROB.
-------------------------------------------------------------------------------------------------------------------------------------
                                                       *                                   *                *
A                                                      *                                   *                *
                        6.6900       .296      .094    *                                   *                *
                                                       *    1.0800       .257      .081    *  .603   .065   *  13.27        9      .000
              10                                       *                                   *                *
                        5.6100       .281      .089    *                                   *                *
                                                       *                                                     
B
-------------------------------------------------------------------------------------------------------------------------------------
```

Table 17. SPSS output for Example 8c.

10.0 Analysis of Variance

10.1 Introduction

Analysis of variance, commonly abbreviated "anova," is a statistical method with which we can test whether two or more sample means are significantly different. This section begins with a discussion of the assumptions underlying the anova tests. Emphasis has been placed on what these assumptions mean and how violations of them affect the validity of the probability statements resulting from the anova tests. This is followed by a discussion, with examples, of several of the more commonly used anova tests: the one-way anova; the two-level nested anova; and the two-way anova. In addition, methods for evaluating the difference between means is also covered.

10.2 Assumptions Underlying Analysis of Variance

There are several assumptions underlying the anova tests. Violations of one or more of these assumptions can affect both the level of significance and sensitivity of the test.

a. Random Sampling

It is assumed that the observations were obtained by random sampling. If our sampling procedure is biased, such as always collecting samples from riffles when we are interested in the substrate composition of an entire stream bed, we are likely to have problems meeting the other assumptions that follow here. If the other assumptions hold, the lack of random sampling will probably have little effect on the level of significance and sensitivity of the anova test. However, the chances of the other assumptions being valid when random sampling was not used are not very good. Consequently, as a precautionary measure, we should consider the concept of random sampling very carefully when we design our monitoring program.

b. Independence

It is assumed that the error terms for each variate are independently and identically distributed. If the error terms are not independent, the validity of the usual F-test of significance can be seriously impaired (Sokal and Rohlf, 1969). In general, if the samples were collected in a random manner, there should be no problem with independence of the error terms. If lack of independence is suspected, such as when evaluating spatial distribution of macroinvertebrates in a stream channel, it can be tested using the "Runs" test (Sokal and Rohlf, 1969). If the errors are not independent, there is no simple way to correct the problem short of redesigning the study.

c. Homogeneity of Variances

The assumption of homogeneity of variances is very important in the anova test because each sample variance is considered to be an estimate of the same parametric error variance. You are likely to encounter the problem of nonhomogeneity of variances when the means of one or two groups are much larger than the others, such as mean $SO_4^=$ concentration in a lake draining an area underlain by granitics as opposed to one draining an area high in gypsum.

Two methods you can use to test for homogeneity of variances between two sample means have already been discussed (Section 8.0). In addition, Bartlett's test can also be used in situations where more than two means are involved. For a quick "first inspection" of this assumption, you can simply check the correlation between the means and variances of the samples. If the variances increase with the means, the coefficient of variation [$s(100)/\bar{X}$] will be approximately constant for each sample. If the means and variances are independent, however, this ratio will vary widely.

If moderate heterogeneity of variances exists, the effect on the overall test of significance is probably not too serious. However, if this condition exists and you are making comparisons with one degree of freedom, you can expect serious problems.

d. Additivity

It is assumed that the treatment (group) and environmental effects are additive. The assumption of additivity can be evaluated using Turkey's test (Sokal and Rohlf, 1969). In general, if this assumption is not met by the data it can be corrected with a data transformation.

Consider the data presented in Table 18. Here we have two fish tanks, A and B, in which the concentration of a heavy metal required to kill 50 percent of the test fish at two different pH levels was determined. Prior to the log (X) transformation, the effects were multiplicative (3 x 10 = 30 and 3 x 20 = 60). However, after the transformation, the effects became additive (1.00 + 0.48 = 1.48 and 1.30 + 0.48 = 1.78).

Table 18. 96-hour LD_{50} concentration (mg/l) for test fish.

	Untransformed		Log transformed	
	pH		pH	
Tank	4.0	6.0	4.0	6.0
A	10	20	1.00	1.30
	(3x)	(3x)	(+0.48)	(+0.48)
B	30	60	1.48	1.78

10.3 Nonparametric Methods of Analysis of Variance

If we cannot transform our data to meet the assumptions of the analysis of variance, we may have to resort to a similar nonparametric method. These methods are not concerned with specific parameters, but only with the distribution of the variates. Although not discussed here, a detailed description of many of these methods is covered by Sokal and Rohlf (1969).

10.4 One-Way Classification Analysis of Variance

The most basic anova test is the one-way anova. Higher order anova tests are merely extensions of the one-way anova. The one-way anova has only one criterion for classification, such as sampling stations, and, in its simplest form, has an equal number of data observations in each group.

The conceptual framework of the one-way classification anova with equal sample sizes is straightforward and easy to follow. Consider "a" sampling stations ("a" groups) along a stream. Suppose we have measured specific conductance "n" times ("n" observations) at each station. Each specific conductance observation can be denoted by X_{ij} which is the jth observation of the ith station. The data may be arranged as in Table 19.

Table 19. Data arranged for a one-way classification anova.

	"a" groups						
	1	2	3	. . .	i	. . .	a
1	X_{11}	X_{21}	X_{31}	. . .	X_{i1}	. . .	X_{a1}
2	X_{12}	X_{22}	X_{32}	. . .	X_{i2}	. . .	x_{a2}
3	X_{13}	X_{23}	X_{33}	. . .	X_{i3}	. . .	X_{a3}
j	X_{1j}	X_{2j}	X_{3j}	. . .	X_{ij}	. . .	$X_a{}^J$
n	X_{1n}	X_{2n}	X_{3n}	. . .	X_{in}	. . .	X_{an}

The null hypothesis that we wish to test can be stated as follows

$$H_o: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_i = \ldots = \mu_a$$

The alternative hypothesis, therefore, is

$$H_a: \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \neq \mu_i \neq \ldots \neq \mu_a$$

The test for differences in means is based on the fact that if the means of each group are greatly different, the variance of the combined groups is much larger than the variances of the separate groups. Consequently, to test for differences in means we test for differences in variances. With the one-way anova we obtain two independent estimates of the population variance; one is based on the variance *within* groups while the other is based on the variances *among* groups.

To test the null hypothesis stated above, the following calculations are required. The *correction term*, C, is determined by Equation 13.

$$C = \frac{1}{an}\left(\sum_{i=1}^{a}\sum_{j=1}^{n} X_{ij}\right)^2 \qquad (13)$$

The *total sum of squares* (SS_{total}), adjusted for the mean, is found using Equation 14.

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n} X_{ij}^2 - C \qquad (14)$$

The sum of squares attributable to groups is commonly called the *between groups sum of squares* (SS_{groups}) or groups sum of squares and calculated using Equation 15.

$$SS_{groups} = \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n} X_{ij}\right)^2 - C \qquad (15)$$

The sum of squares within a group is referred to as the *within group sum of squares* (SS_{within}), as well as residual sum of squares and/or error sum of squares, and is generally found by subtracting the between groups sum of squares from the total sum of squares, Equation 16. This can be done because the sums of squares are additive.

$$SS_{within} = SS_{total} - SS_{groups} \qquad (16)$$

The results of a one-way classification anova with equal sample sizes are generally presented in an anova table similar to Table 20. Table 20 is divided into five columns. Column (1) identifies the source of variation as among groups, within groups and total. Column (2) gives the degrees of freedom by which the various sums of squares must be divided in order to yield estimates of the variances. Column (3) lists the sums of squares (SS) for the respective sources of variation. Column (4) contains the mean square (MS). The mean square is obtained by dividing the sum of squares by the degrees of freedom. The two mean squares obtained are the two estimates of the variance discussed earlier. Column 5 presents the calculated F statistic, F_s. It is defined as the ratio of the two independent estimates of the same population variance. The mean square obtained from the means (between groups) is always placed in the numerator since we wish to state that the means are significantly different only if they are significantly more spread out than would be expected for samples from the same population.

Two examples of the one-way classification anova with equal sample size are presented in Examples 9a and 9b.

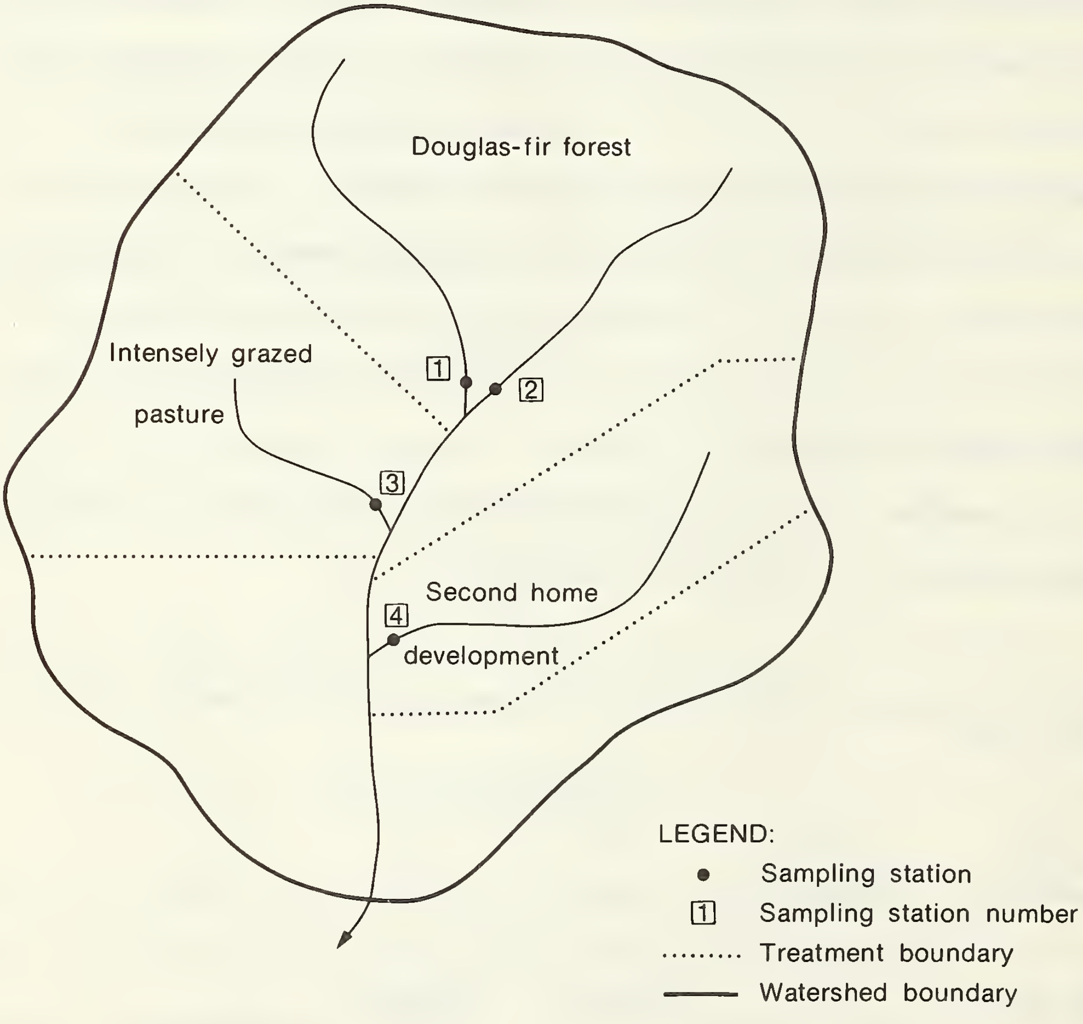
Table 20. A typical ANOVA table for the one-way classification with equal sample sizes.

(1) Source of Variation	(2) Degrees of Freedom	(3) SS	(4) MS	(5) F_s
among groups	a-1	SS_{groups}	$\frac{SS_{groups}}{(a-1)}$	$\frac{MS_{groups}}{MS_{within}}$
within groups	a(n-1)	SS_{within}	$\frac{SS_{within}}{a(n-1)}$	
total	an-1	SS_{total}		

EXAMPLE 9a
One-Way Classification ANOVA with Equal Sample Size

Problem:

Nitrate concentrations were monitored at four stations within a watershed over a one-year period (see illustration below). The results have been tabulated below the watershed illustration. Determine if there is a significant difference, at the 5% level, in the mean NO_3 concentration between stations.



NITRATE CONCENTRATIONS (mg/l)

	STATIONS			
	1	2	3	4
	8.7	9.8	11.9	8.8
	8.0	8.6	15.1	7.9
	8.9	8.8	11.2	8.5
	8.0	8.3	11.9	8.1
	6.8	7.4	13.9	10.0
	6.4	9.4	13.7	7.6
	7.8	7.9	12.6	10.1
	8.4	8.9	16.3	9.2
	7.8	8.3	15.4	10.0
	7.7	11.1	14.4	8.5
	8.3	8.9	13.2	12.7
	8.3	8.2	11.8	9.6
	9.7	10.7	12.6	8.5
	6.9	7.2	12.1	10.2
	7.4	7.2	13.3	6.6
ΣX	119.10	130.70	199.40	136.30
$\bar{X}$	7.94	8.71	13.29	9.09
s	0.85	1.16	1.50	1.44

Solution:

1. Establish the hypothesis to be tested.

$$H_o: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

2. Selected the significance level.

From the problem statement, $\alpha = 0.05$.

3. Develop the anova table and determine the test statistic, F_S.

The data were analyzed using the ONEWAY subprogram from SPSS (1975). It should be noted that either the ANOVA procedure from SAS (1979) or the ANOVA subprogram from SPSS (1975) could have been used to solve for the test statistic. The results of the analysis are presented in Table 21. From Table 21 it can be seen that $F_S = 54.133$.

4. Define the critical region, F_c.

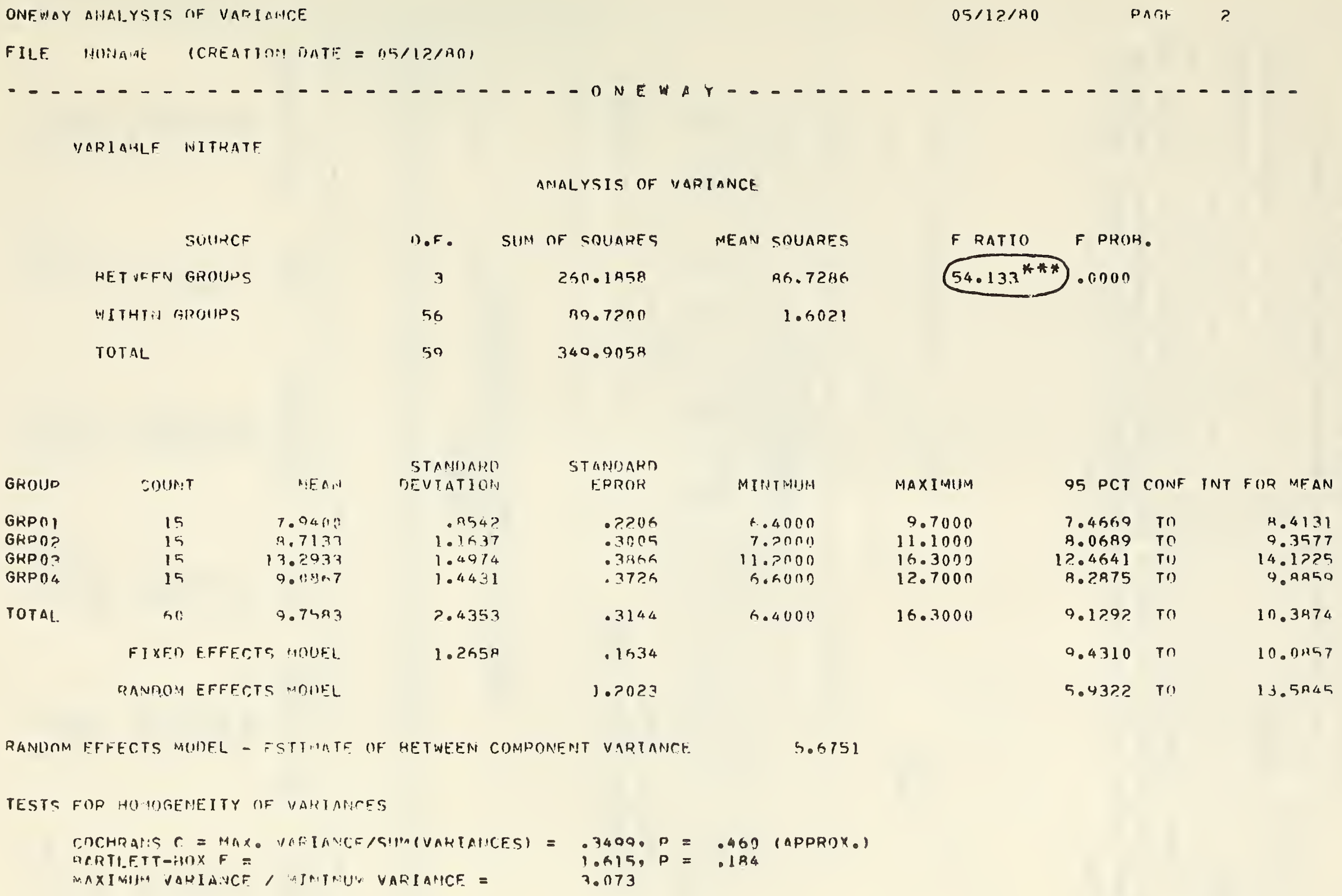
From Table A-4 we find that

$$F_c = F_{.05(3.56)} \simeq 2.76$$

5. Reject or accept the null hypothesis.

Since $F_s > F_c$, we do not accept the null hypothesis. This means that the mean NO_3^- concentrations at the four stations cannot be considered equal (i.e. having come from the same population) with our $\alpha = 0.05$.

It should be noted that at this point we have no statistical insight as to which mean or means differ from each other. All we know is that they are not all equal.



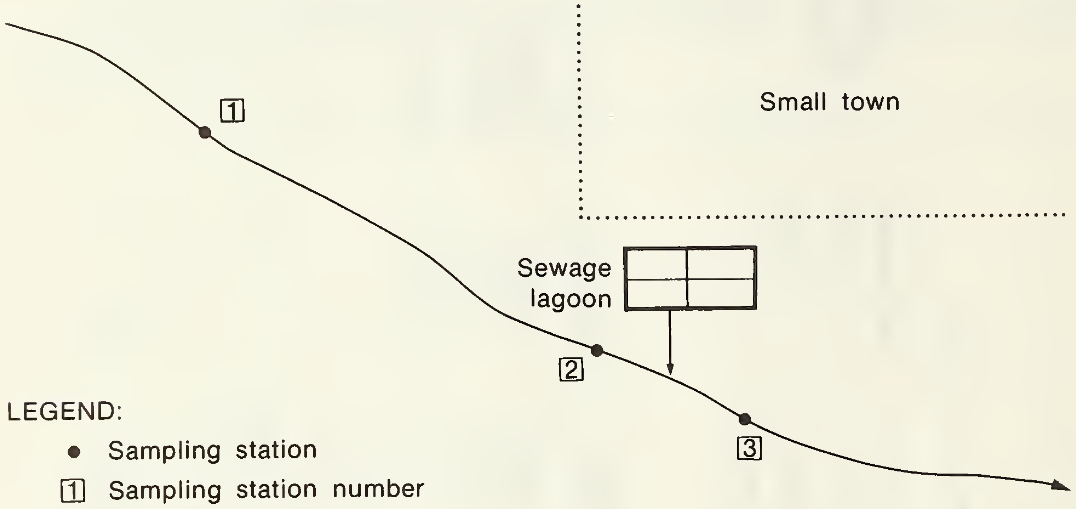
```
ONEWAY ANALYSIS OF VARIANCE                                                          05/12/80        PAGE    2

FILE   NONAME    (CREATION DATE = 05/12/80)

- - - - - - - - - - - - - - - - - - - - - - - - - - - O N E W A Y - - - - - - - - - - - - - - - - - - - - - - - - - - -

      VARIABLE   NITRATE

                                                  ANALYSIS OF VARIANCE

                SOURCE              D.F.     SUM OF SQUARES     MEAN SQUARES        F RATIO    F PROB.

        BETWEEN GROUPS                 3         260.1858          86.7286          54.133***   .0000

        WITHIN GROUPS                 56          89.7200           1.6021

        TOTAL                         59         349.9058




                                        STANDARD       STANDARD
GROUP          COUNT         MEAN      DEVIATION        ERROR          MINIMUM       MAXIMUM         95 PCT CONF INT FOR MEAN

GRP01            15         7.9400        .8542          .2206          6.4000        9.7000         7.4669  TO       8.4131
GRP02            15         8.7133       1.1637          .3005          7.2000       11.1000         8.0689  TO       9.3577
GRP03            15        13.2933       1.4974          .3866         11.2000       16.3000        12.4641  TO      14.1225
GRP04            15         9.0867       1.4431          .3726          6.6000       12.7000         8.2875  TO       9.8859

TOTAL            60         9.7583       2.4353          .3144          6.4000       16.3000         9.1292  TO      10.3874

             FIXED EFFECTS MODEL         1.2658          .1634                                       9.4310  TO      10.0857

            RANDOM EFFECTS MODEL                        1.2023                                       5.9322  TO      13.5845

RANDOM EFFECTS MODEL - ESTIMATE OF BETWEEN COMPONENT VARIANCE          5.6751

TESTS FOR HOMOGENEITY OF VARIANCES

      COCHRANS C = MAX. VARIANCE/SUM(VARIANCES) =   .3499, P =  .460 (APPROX.)
      BARTLETT-BOX F =                             1.615, P =  .184
      MAXIMUM VARIANCE / MINIMUM VARIANCE =        3.073
```

Table 21. Results of the Analysis of Variance for Example 9a using the ONEWAY subprogram from SPSS (1975).

EXAMPLE 9b
One-Way Classification ANOVA with Equal Sample Size and Data Transformed by Log (X)

Problem:

Stonefly Nymphs were collected at three stations along a stream (see illustration below). The sample results are tabulated below. Determine if the mean counts are significantly different (P = 0.05) between sampling stations.



RESULTS OF STONEFLY NYMPH SAMPLES

STATIONS

	1		2		3	
	Counts	log (counts)	Counts	log (Counts)	Counts	log (Counts)
	91	1.96	120	2.08	8	0.90
	77	1.89	110	2.04	17	1.23
	86	1.93	93	1.97	20	1.30
	52	1.72	150	2.18	15	1.18
	80	1.90	82	1.91	10	1.00
ΣX	386	9.40	555	10.18	70	5.61
$\bar{X}$	77.2	1.88	111.00	2.04	14.00	1.12
s	15.09	0.09	26.31	0.10	4.95	0.17

Solution:

A common problem with count data, such as that presented here, is that the variance of the sample is not independent of the mean. Generally, this can be corrected with a log (X) or log (X + 1) transformation. This transformation does not affect the validity of the anova and, in fact in this case, makes the results more reliable. The calculations are carried out in the same manner as with the nontransformed data.

The solution procedure is as follows.

1. Establish the hypothesis to be tested.

$$H_o: \mu_1 = \mu_2 = \mu_3$$

$$H_a: \mu_1 \neq \mu_2 \neq \mu_3$$

2. Select the significance level.

From the problem statement, $\alpha = 0.05$.

3. Develop the anova table and determine the test statistic, F_S.

The data were analyzed using the same procedures outlined in step 3 of the solution of Example 9a. However, in this case, the log-transformed data were input into the program rather than the raw data. The results of the analysis are presented in Table 22. From Table 22, it can be seen that $F_S = 75.97$.

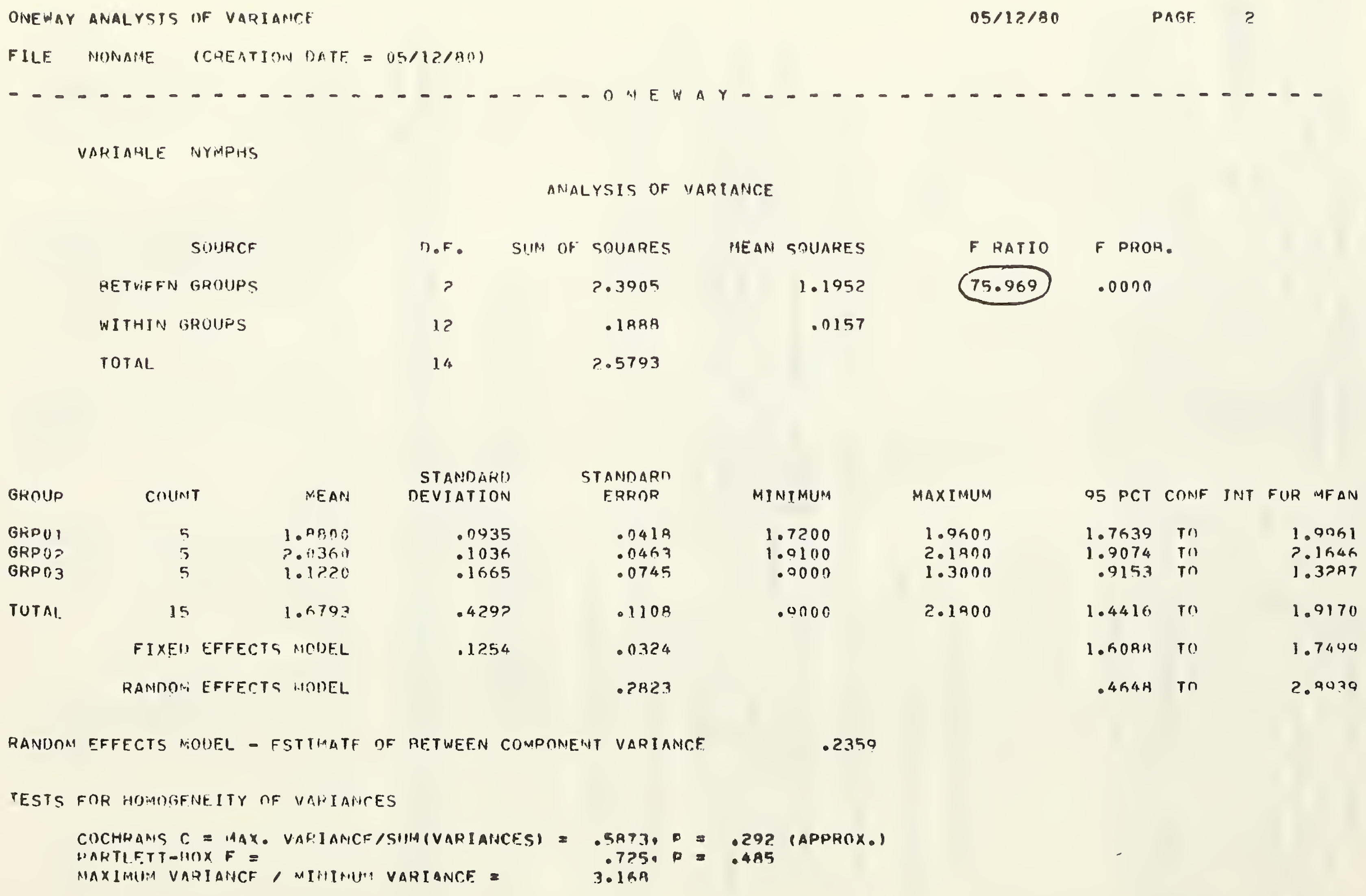
4. Define the critical region, F_C.

From Table A-4 we find that

$$F_C = F_{.05(2,12)} = 3.88$$

5. Reject or accept the null hypothesis.

Since $F_S > F_C$, we do not accept the null hypothesis. This means that the mean Stonefly Nymph counts at the three stations are not equal for $\alpha = 0.05$.



```
ONEWAY ANALYSIS OF VARIANCE                                                   05/12/80        PAGE    2

FILE    NONAME    (CREATION DATE = 05/12/80)

- - - - - - - - - - - - - - - - - - - - - - - - - O N E W A Y - - - - - - - - - - - - - - - - - - - - - - - - -

      VARIABLE  NYMPHS
                                              ANALYSIS OF VARIANCE

             SOURCE               D.F.    SUM OF SQUARES    MEAN SQUARES         F RATIO    F PROB.

        BETWEEN GROUPS              2          2.3905           1.1952            75.969      .0000

        WITHIN GROUPS              12           .1888            .0157

        TOTAL                      14          2.5793



                                          STANDARD       STANDARD
GROUP       COUNT          MEAN          DEVIATION        ERROR         MINIMUM       MAXIMUM        95 PCT CONF INT FOR MEAN

GRP01          5         1.8800            .0935          .0418         1.7200        1.9600         1.7639   TO        1.9961
GRP02          5         2.0360            .1036          .0463         1.9100        2.1800         1.9074   TO        2.1646
GRP03          5         1.1220            .1665          .0745          .9000        1.3000          .9153   TO        1.3287

TOTAL         15         1.6793            .4292          .1108          .9000        2.1800         1.4416   TO        1.9170

            FIXED EFFECTS MODEL            .1254          .0324                                      1.6088   TO        1.7499

           RANDOM EFFECTS MODEL                           .2823                                       .4648   TO        2.8939

RANDOM EFFECTS MODEL - ESTIMATE OF BETWEEN COMPONENT VARIANCE           .2359

TESTS FOR HOMOGENEITY OF VARIANCES

      COCHRANS C = MAX. VARIANCE/SUM(VARIANCES) =  .5873, P =  .292 (APPROX.)
      BARTLETT-BOX F =                              .725, P =  .485
      MAXIMUM VARIANCE / MINIMUM VARIANCE =        3.168
```

Table 22. Results of the Analysis of Variance for Example 9b using the ONEWAY subprogram from SPSS (1975).

In the case when groups are not composed of samples of equal size, the required computations for the one-way classification anova are as follows.

Correction term, C.

$$C = \left(\sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}\right)^2 \Big/ \sum_{i=1}^{a} n_i \tag{17}$$

Total sum of squares adjusted for the mean, SS_{total}.

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n} X_{ij}^2 - C \tag{18}$$

Between groups sum of squares, SS_{groups}.

$$SS_{groups} = \sum_{j=1}^{a} \frac{\left(\sum_{j=1}^{n} X_{ij}\right)^2}{n_i} - C \tag{19}$$

Within groups sum of squares, SS_{within}.

$$SS_{within} = SS_{total} - SS_{groups} \tag{20}$$

The results of a one-way anova with unequal sample sizes are generally presented in an anova table similar to Table 23.

Table 23. A typical ANOVA table for the one-way classification with unequal sample size.

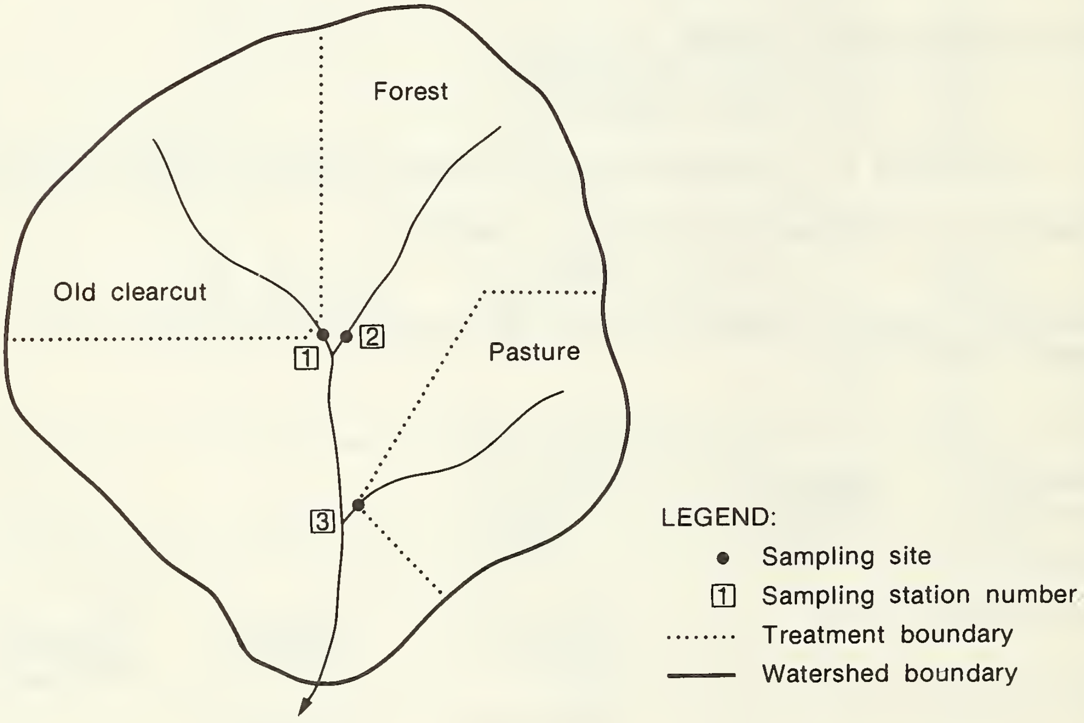
Source of Variation	Degrees of Freedom	SS	MS	FS
among groups	a-1	SS_{groups}	$\frac{SS_{groups}}{a-1}$	$\frac{MS_{groups}}{MS_{within}}$
within groups	n-a	SS_{within}	$\frac{SS_{within}}{n-a}$	
total	n-1	SS_{total}		

An example of the one-way classification anova with unequal sample sizes within groups is presented in Example 9c.

EXAMPLE 9c
One-Way Classification ANOVA with Unequal Sample Size

Problem:

Suspended solids concentration was monitored at three stations within a watershed (see illustration). The results are tabulated below the illustration. Determine if there is a significant difference (P = 0.10) in the mean SS concentration between stations.



STATION

	1			2			3	
	mg/l	log (X)		mg/l	log (X)		mg/l	log (X)
	27	1.43		49	1.69		17	1.23
	29	1.46		25	1.40		30	1.48
	24	1.38		13	1.11		25	1.40
	48	1.68		29	1.46		18	1.26
	69	1.84		46	1.66		23	1.36
	30	1.48		15	1.18		18	1.26
	21	1.32		29	1.46		17	1.23
	68	1.83		23	1.36		10	1.00
	21	1.32		15	1.18		20	1.30
	20	1.30		28	1.45		21	1.32
	30	1.48	ΣX	272	13.95		37	1.57
	74	1.87	$\bar{X}$	27.2	1.40	ΣX	236	14.40
	26	1.41	s	12.3	0.20	$\bar{X}$	21.4	1.31
ΣX	487	19.80				s	7.3	0.15
$\bar{X}$	37.5	1.52						
s	20.1	0.21						

Solution:

It is obvious that initially the sample variances and means do not vary independently. However, after the log (X) transformation, this problem was corrected.

1. Establish the hypothesis to be tested.

H_o: $\mu_1 = \mu_2 = \mu_3$

H_a: at least one mean is not equal to the other means

2. Select the significance level.

From the problem statement, $\alpha = 0.10$.

3. Develop the anova table and determine the test statistic, F_S.

The log-transformed data were analyzed using the ANOVA procedure from SAS (1979). It should be noted that either the ONEWAY or ANOVA subprograms from SPSS (1975) could have been used to solve for the test statistic. The results of the analysis are presented in Table 24 where it can be seen that $F_S = 3.95$.

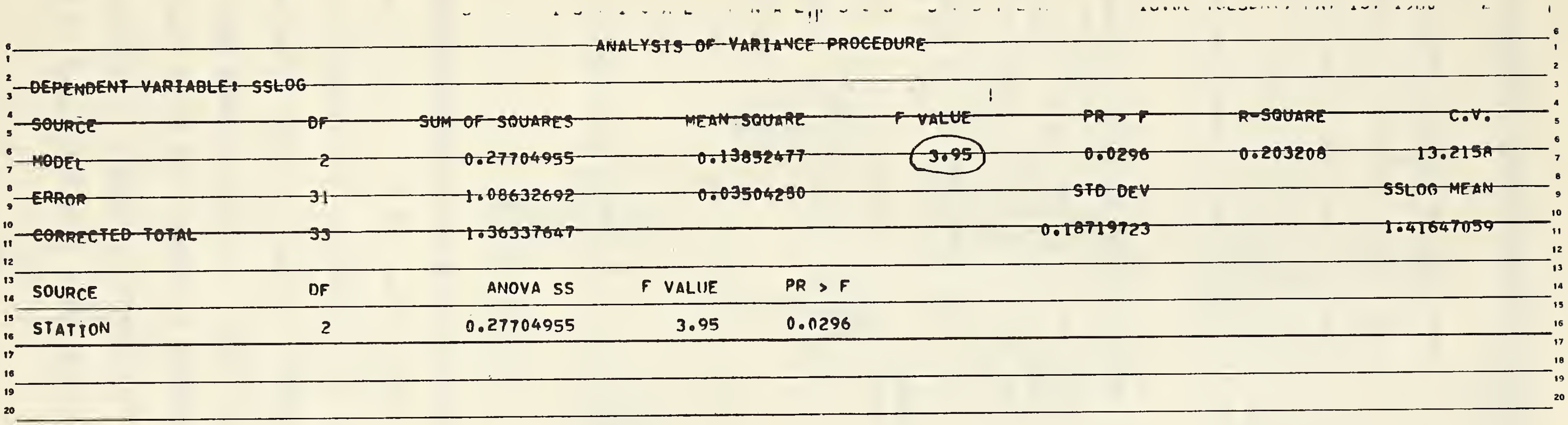
4. Define the critical region, F_c.

From Table A-4 we find that

$$F_c = F_{.05(2,31)} = 2.49$$

5. Accept or reject the null hypothesis.

Since $F_s > F_c$ we do not accept the null hypothesis.



```
                                    ANALYSIS OF VARIANCE PROCEDURE

DEPENDENT VARIABLE: SSLOG

SOURCE              DF     SUM OF SQUARES      MEAN SQUARE       F VALUE       PR > F      R-SQUARE          C.V.

MODEL                2         0.27704955       0.13852477          3.95       0.0296      0.203208       13.2158

ERROR               31         1.08632692       0.03504280                    STD DEV                  SSLOG MEAN

CORRECTED TOTAL     33         1.36337647                                  0.18719723                  1.41647059


SOURCE              DF           ANOVA SS     F VALUE      PR > F

STATION              2         0.27704955        3.95      0.0296
```

Table 24. Results of the analysis of variance for Example 9c using the ANOVA procedure from SAS (1979).

10.5 Evaluating the Difference Between Means When the Sample Sizes are Equal

If we reject the null hypothesis after performing the anova test, all we know is that there is a significant difference (at the α level selected) between treatment means. We have no idea which group (or groups) is responsible for our rejection of the null hypothesis. In some cases we may want to isolate the group (or groups) which causes us to reject the null hypothesis.

A simple method of evaluating the difference between means has been developed by Snedecor (1956) and is called the Q-test. (Much of what follows has been taken from Nash, 1965.) The procedure is outlined using a hypothetical data set. Initially, the means of the groups are ranked in order from highest to lowest.

Order	Group	Mean
1	C	5.3
2	A	4.8
3	B	4.4
4	D	4.0

Next, we determine the difference, D, using Equation 21, for each value of Q.

$$D = Qs_{\bar{x}} = Q\sqrt{\frac{s^2}{n}} \tag{21}$$

where s^2 is the mean square of the error or within groups term, n is the number of observations in each group and Q is a factor obtained from Table A-5 which gives the upper five percent of the range for different degrees of freedom and for the number of groups. To determine which group (or groups) is causing a significant difference between means, we use a different value of Q depending on whether the group means are 2, 3, . . ., or n ranks apart. In our hypothetical example, assume we have 12 degrees of freedom associated with the error term (SS_{within}), a MS_{within} of

0.1923 and four observations per group. Now, we can determine D for each value of Q.

$$s_{\bar{x}} = \sqrt{\frac{0.1923}{4}} = 0.2193$$

$$a = 2 \quad D = 3.08(0.2193) = 0.6754$$

$$a = 3 \quad D = 3.77(0.2193) = 0.8268$$

$$a = 4 \quad D = 4.20(0.2193) = 0.9211$$

Next, we rank the means in order and determine the differences between the highest and lowest. For simplicity, and to demonstrate the differences between means and the comparisons with D, the values of D have been inserted in parentheses. If the difference between means is greater than the value of D for a particular difference, the difference between means is significant at the 5% level. For example, X - 4.0 = 1.3 is greater than D = 0.9211; therefore the difference between the means of Group C and D is significant. However, the difference between the means of groups A and D is not significant, and so on.

Group	Mean	X - 4.0	X - 4.4	x - 4.8
C	5.3	1.3 (0.9211)	0.9 (0.8268)	0.5 (0.6754)
A	4.8	0.8 (0.8268)	0.4 (0.6754)	-
B	4.4	0.4 (0.6754)	-	
D	4.0	-		

An example of the Snedecor Q-test is presented in Example 9d.

EXAMPLE 9d
Evaluating the Differences Between Means Using Snedecor's Q-test

---

Problem:

When we evaluated the NO_3 data presented in Example 9a, it was determined that there was a significant difference, at the 1% level, in the mean NO_3 concentrations between stations. Now, evaluate the differences between the means and determine which ones differ significantly (P = 0.05) using Snedecor's Q-test.

Solution:

1. Rank the means in order from highest to lowest.

Rank	Station	Mean
1	3	13.29
2	4	9.09
3	2	8.71
4	1	7.91

2. Determine the difference, D, using Equation 21 for each value of Q. Q is obtained from Table A-5.

$$D = Qs_{\bar{x}}$$

where

$$s_{\bar{x}} = \frac{1.60}{15} = 0.3266$$

a = 2 $D = 2.83(0.3266)$
a = 3 $D = 3.40(0.3266)$
a = 4 $D = 3.74(0.3266)$

Note: Since Table A-5 does not list 56 df, 60 was used to obtain Q. The difference in Q between 40 and 60 df is very minor.

3. Rank the means in order and determine the differences between the highest and lowest.

Rank	Station	Mean	X - 7.94	X - 8.71	X - 9.09
1	3	13.29	5.35 (1.22)**	4.58 (1.11)**	4.20 (0.92)**
2	4	9.09	1.15 (1.11)**	0.38 (0.92)**	-
3	2	8.71	0.77 (0.92)	-	-
4	1	7.94	-	-	-

4. Interpret the results.

As you recall, if the difference between means is greater than the value of D for a particular difference, the difference between means is significant at the 5% level. The results of this analysis indicate the mean at Station 3 is significantly different from the means at all the other stations while the mean at Station 4 is significantly different from the means at Stations 1 and 3. In other words, the intensely grazed pasture is yielding an average NO_3 concentration significantly greater than any of the other treatments, while the second home development is significantly different from the pasture and one of the forest stations at the 5% level.

10.6 Evaluating the Difference Between Means When the Sample Sizes are Not Equal

In some cases you may want to evaluate the difference between means when the sample sizes are not equal. Although this cannot be done easily by hand, it can be performed very readily on SAS using PROC DUNCAN or on SPSS using the Duncan option with subprogram ONEWAY, both of which are the Duncan's multiple range test. These methods are summarized in SAS (1979) and SPSS (1975).

10.7 Two-Level Nested Analysis of Variance

In studies where we only take a single water quality measurement at a station per visit, we can never be certain that the observed differences in water quality between stations are due solely to environmental or treatment factors alone. There may also be some experimental error as well. The only way to separate these two effects is to take two or more measurements (replicate samples) at a station per visit. If we do not find any significant differences among the replicate samples at a station, we can then ascribe the differences among the stations to environmental or treatment factors.

In the two-level nested anova each group is subdivided into randomly chosen subgroups. Consider the situation where we have three sampling stations, 1, 2, and 3, on a stream where we are studying dissolved solids concentration. Each time we sampled at a station, we obtained three replicate water samples (these samples represent our randomly chosen subgroups) for total dissolved solids determination. The data are symbolized in Table 25. Each dissolved solids determination is denoted by X_{ijk} where i represents the group or station (i = 1, 2, . . . , a), j the

Table 25. Data arranged for a two-level nested anova.

	Station or Groups (a=3)											
	1				2				3			
	Subgroups (b=4)											
	1	2	3	4	1	2	3	4	1	2	3	4
Samples 1	X_{111}	X_{112}	X_{113}	X_{114}	X_{211}	X_{212}	X_{213}	X_{214}	X_{311}	X_{312}	X_{313}	X_{314}
2	X_{121}	X_{122}	X_{123}	X_{124}	X_{221}	X_{222}	X_{223}	X_{224}	X_{321}	X_{322}	X_{323}	X_{324}
3	X_{131}	X_{132}	X_{133}	X_{134}	X_{231}	X_{232}	X_{233}	X_{234}	X_{331}	X_{332}	X_{333}	X_{334}
4	X_{141}	X_{142}	X_{143}	X_{144}	X_{241}	X_{242}	X_{243}	X_{244}	X_{341}	X_{342}	X_{343}	X_{344}
5	X_{151}	X_{152}	X_{153}	X_{154}	X_{251}	X_{252}	X_{253}	X_{254}	X_{351}	X_{352}	X_{353}	X_{354}

sample number (j = 1, 2, . . . , n), and k the randomly chosen subgroup (k = 1,2 . . . , b).

The equations required for the two-level nested anova with equal sample size are as follows:

Correction term, C.

$$C = \left(\sum_{i=1}^{a}\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}\right)^2 \Big/ nba \qquad (22)$$

Total sum of squares, SS_{total}.

$$SS_{total} = \sum_{i=1}^{a}\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}^2 - C \qquad (23)$$

Between groups sum of squares, SS_{groups}.

$$SS_{groups} = \left[\frac{\sum_{i=1}^{a}\left(\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}\right)^2}{nb}\right] - C \qquad (24)$$

Subgroups within groups sum of squares, SS_{subgr}.

$$SS_{subgr} = \left[\frac{\sum_{i=1}^{a}\sum_{k=1}^{b}\left(\sum_{j=1}^{n} X_{ikj}\right)^2}{n}\right] - \left[\frac{\sum_{i=1}^{a}\left(\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}\right)^2}{nb}\right] \qquad (25)$$

Within groups sum of squares, SS_{within}.

$$SS_{within} = SS_{total} - \left[\frac{\sum_{i=1}^{a}\sum_{k=1}^{b}\left(\sum_{j=1}^{n} X_{ikj}\right)^2}{n}\right] \qquad (26)$$

The anova table for the two-level nested anova is illustrated in Table 26.

Table 26. The anova table for the two-level nested anova.

Source of Variation	Degrees of Freedom	SS	MS	F_s
among groups	a-1	SS_{groups}	$\frac{SS_{groups}}{a-1}$	$\frac{MS_{groups}}{MSsubgr}$
among subgroups within groups	a(b-1)	SS_{subgr}	$\frac{SSsubgr}{a(b-1)}$	$\frac{MS_{subgr}}{MS_{within}}$
within subgroups	ab(n-1)	SS_{within}	$\frac{SS_{within}}{ab(n-1)}$	
total	abn-1	SS_{total}		

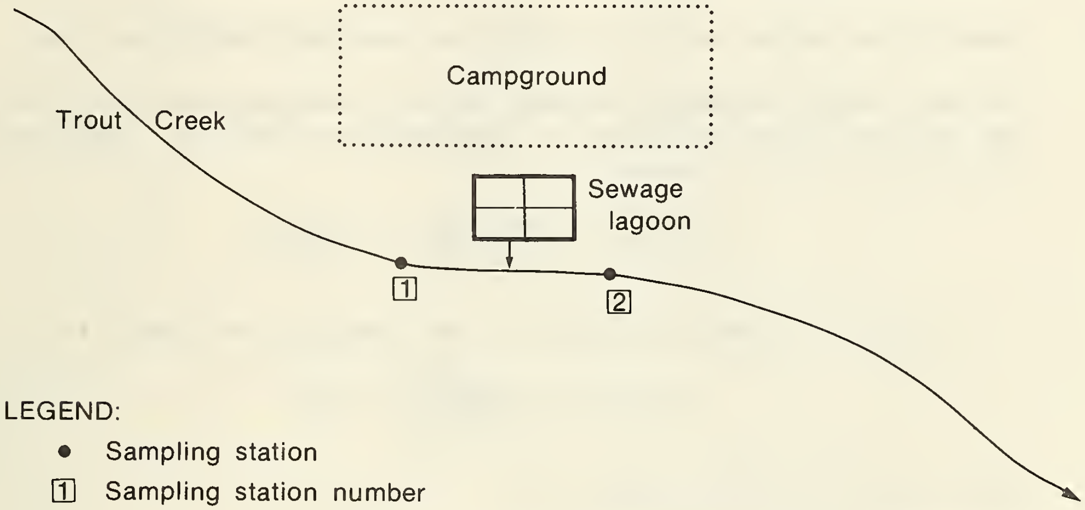
In the case of unequal sample sizes, the required computations for the two-level nested analysis of variance are somewhat different. The equations to use can be readily found in any good statistics book, such as Sokal and Rohlf (1969).

An example of the two-level nested anova is presented in Example 9e.

EXAMPLE 9e
Two-Level Nested ANOVA

Problem:

Nitrate concentration was monitored at two stations along Trout Creek. One station was located above and the other below the outfall from a sewage lagoon servicing a campground (see illustration below). The objective of the study was to determine the onsite effect of the outfall on the nitrate concentration of Trout Creek. At the outset of the study there was some concern over the precision of the analytical method as well as the sample collection procedures. As a result, it was decided that the samples would be collected in replicate in order to separate the treatment effect from the experimental error. The sample results are presented below. Determine if there is a significant difference (P = 0.05) between stations and within stations.



Nitrate Concentrations (mg/l)

Above			Below		
Replicate sample number			Replicate sample number		
1	2	3	1	2	3
1.0	1.1	1.1	5.1	5.3	5.1
1.6	1.5	1.6	6.0	5.8	6.1
1.3	1.3	1.3	5.8	5.9	5.9
1.4	1.3	1.3	6.5	6.5	6.4
1.5	1.5	1.6	6.7	6.6	6.8
2.0	1.8	1.9	6.1	6.1	6.1
2.1	2.0	2.0	6.9	6.8	6.9
1.7	1.8	1.7	5.5	5.5	5.6
1.6	1.6	1.6	5.4	5.4	5.4

Solution:

1. Establish the hypotheses to be tested.

   a. Site (stations).

   $H_o: \mu_A = \mu_B$

   $H_a: \mu_A \neq \mu_B$

   b. Experimental error.

   $H_o: \mu_1 = \mu_2 = \mu_3$

   $H_a: \mu_1 \neq \mu_2 \neq \mu_3$

2. Select the level of significance.

   From the problem statement, $\alpha = 0.05$.

3. Develop the anova table and determine the test statistics, F_s.

   The data were analyzed using the NESTED procedure from SAS (1979). The results of the analysis are presented in Table 27. The test statistic for the stations (site) is

$$F_{s(site)} = \frac{266.66667}{0.22181} = 1202.25$$

   while the test statistic for the replications (num) is

$$F_{s(num)} = \frac{0.00370}{0.22181} = 0.02$$

4. Define the critical regions, F_c.

   From Table A-4 we find that

   a. Site

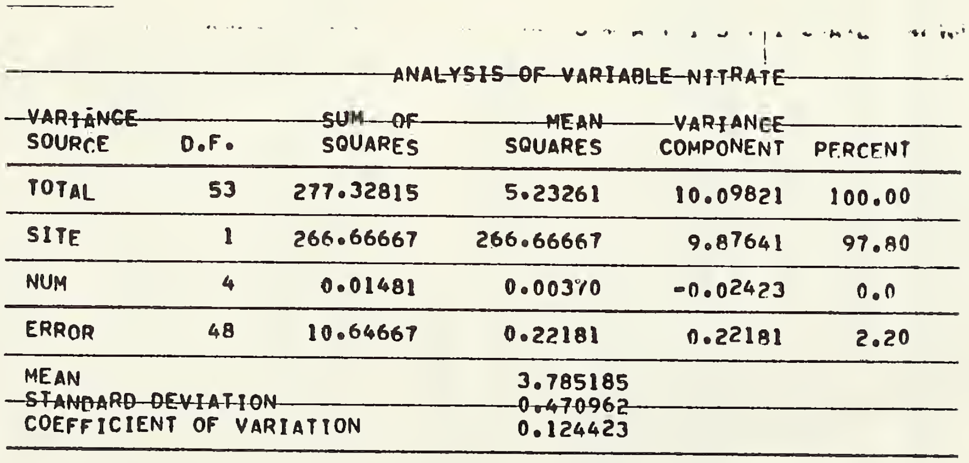
   $F_{c(site)} = F_{.05(1,43)} = 4.04$

   b. Replications

   $F_{c(num)} = F_{.05(4,48)} = 2.56$

5. Accept or reject the null hypothesis.

Since $F_{s(site)} > F_{c(site)}$, we reject the null hypothesis that the mean NO_3^- concentrations between sites are equal. However, since $F_{s(num)} < F_{c(num)}$, we do not reject the null hypothesis stating the mean NO_3 concentrations within samples are equal.



ANALYSIS OF VARIABLE NITRATE

VARIANCE SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	VARIANCE COMPONENT	PERCENT
TOTAL	53	277.32815	5.23261	10.09821	100.00
SITE	1	266.66667	266.66667	9.87641	97.80
NUM	4	0.01481	0.00370	-0.02423	0.0
ERROR	48	10.64667	0.22181	0.22181	2.20
MEAN			3.785185		
STANDARD DEVIATION			0.470962		
COEFFICIENT OF VARIATION			0.124423		

Table 27. Results of the analysis of variance for Example 9e using the NESTED procedure from SAS (1979).

10.8 Two-Way Classification Analysis of Variance

The two-way classification anova allows us to evaluate the effects of two factors, such as stations and seasons, simultaneously. It is assumed in this method that each factor contributes to the water quality and that the two factors add their effects without influencing each other.

Care should be taken in the design of your data analysis so that you do not confuse a two-level nested anova with a two-way anova. Consider Table 28. Here we have several sampling stations for which we have replicate samples, denoted by 1 and 2. Replicate samples 1 and 2 are simply arbitrary designations for two randomly selected samples at each station. Replicate sample 1 at station 1 has no closer relation to replicate sample 1 at station 2 than it does to replicate sample 2 station 1.

Table 28. Basic design of the two-level nested anova.

	Stations								
	1		2		3		. . .	a	
Replicate samples	1	2	1	2	1	2		1	2

Now, consider Table 29 an example of a two-way anova. Here we have several sampling stations for which we have collected samples during two seasons, spring and fall. Why can we not rearrange the seasons into a nested design? The reason we cannot do this is because the seasons are common to all sites within the study. If we nested the seasons at each station, this would imply that the two seasons per station were random samples from all possible seasons and that spring at Station 1 is not the same as spring at Station 2.

Table 29. Basic design of a two-way anova.

Station

	1		2		3		. . .	a	
Spring									
Fall									

Sokal and Rohlf (1969) state that the critical question to be asked is always, "Does the arrangement of the data into a two-way table falsely imply a correspondence across classes?" If it does, and we recognize that the factor represents only random subdivisions of the groups of another factor, then we have a nested anova. If there is correspondence across groups, the two-way design is appropriate.

Although there are many different designs of a two-way anova, the one you are most likely to need is the two-way anova with replicate samples. The basic design of the two-way anova with replicate samples is illustrated in Table 30. The data are classified two ways, by station (column) and by season (row). In this test we want to test for differences in the means among stations and among seasons, and assess the interaction of the two factors.

The equations required for the two-level anova with replicate samples are as follows.

Correction term, C.

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}\right)^2}{abn} \qquad (27)$$

Total sum of squares, SS_{total}.

$$SS_{total} = \sum_{i=1}^{a}\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}^2 - C \qquad (28)$$

Table 30. Total dissolved solids concentration (mg/l) at Stations 001 and 002 on Eagle Creek during the spring and fall of 1980.

	Station (c = 2)	
Season (r = 2)	001	002
Spring	340 390 381	512 560 550
Fall	612 630 633	917 920 915

Subgroups within groups sum of squares, SS_{subgr}.

$$SS_{subgr} = \frac{\sum_{i=1}^{a}\sum_{k=1}^{b}\left(\sum_{j=1}^{n} X_{ikj}\right)^2}{n} - C \qquad (29)$$

Sum of squares of columns, SS_A.

$$SS_A = \frac{\sum_{i=1}^{a}\left(\sum_{k=1}^{b}\sum_{j=1}^{n} X_{ikj}\right)^2}{bn} - C \qquad (30)$$

Sum of squares of rows, SS_B.

$$SS_B = \frac{\sum_{k=1}^{b}\left(\sum_{i=1}^{a}\sum_{j=1}^{n} X_{ikj}\right)^2}{bn} - C \qquad (31)$$

Interaction sum of squares, $SS_{A \times B}$.

$$SS_{AxB} = SS_{subgr} - SS_A - SS_B \qquad (32)$$

Within subgroups sum of squares, SS_{within}.

$$SS_{within} = SS_{total} - SS_{subgr} \qquad (33)$$

The anova table for the two-way anova with replicate samples is illustrated in Table 31.

Table 31. The anova table for the two-way anova with replicate samples.

Source of variation	Degrees of freedom	SS	MS	F_s
Subgroups	ab-1	SS_{subgr}	$\frac{SS_{subgr}}{ab-1}$	
A (columns)	a-1	SS_A	$\frac{SS_A}{a-1}$	$\frac{MS_A}{MS_{within}}$
B (rows)	b-1	SS_B	$\frac{SS_B}{b-1}$	$\frac{MS_B}{MS_{within}}$
A X B (interaction)	(a-1)(b-1)	$SS_{A\ X\ B}$	$\frac{SS_{A\ X\ B}}{(a-1)(b-1)}$	$\frac{MS_{A\ X\ B}}{MS_{within}}$
Within Subgroups	ab(n-1)	SS_{within}	$\frac{SS_{within}}{ab(n-1)}$	
Total	abn-1	SS_{total}		

An example of the two-way anova with replication is presented in Example 9f.

EXAMPLE 9f
Two-Way Classification ANOVA

Problem:

Suspended solids concentration was monitored at the mouths of two adjacent watersheds, A and B, over a period of one year. Watershed A has been 30% clearcut while Watershed B is fully vegetated. The results are tabulated below. Determine if there is a significant difference ($\alpha = 0.01$) in the mean SS concentration between watersheds (1) on an annual basis and (2) by season (spring-summer vs. fall-winter).

Suspended Solids Concentration

	Watershed	
Season	A	B
Spring-Summer	60 75 83 69 58 89	27 22 25 24 26 29
Fall-Winter	57 45 59 61 38 40	20 21 17 15 18 19

Solution:

1. Establish the hypothesis to be tested:

   a. Stations

   H_o: $\mu_A = \mu_B$

   H_a: $\mu_A \neq \mu_B$

   b. Seasons

   H_o: $\mu_{SS} = \mu_{FW}$

   H_a: $\mu_{SS} \neq \mu_{FW}$

2. Select the significance level.

   From the problem statement, $\alpha = 0.05$.

3. Develop the anova table and determine the test statistics, F_s.

   The data were analyzed using the ANOVA subprogram from SPSS (1975). It should be noted that the ANOVA procedure from SAS could have been used to solve for the test statistics. The results of the analysis are presented in Table 32. From Table 32 it can be seen that the test statistic for the seasons is equal to 19.481, the test statistic for the stations is 137.944 and the test statistic for the interaction between station and season is 5.149.

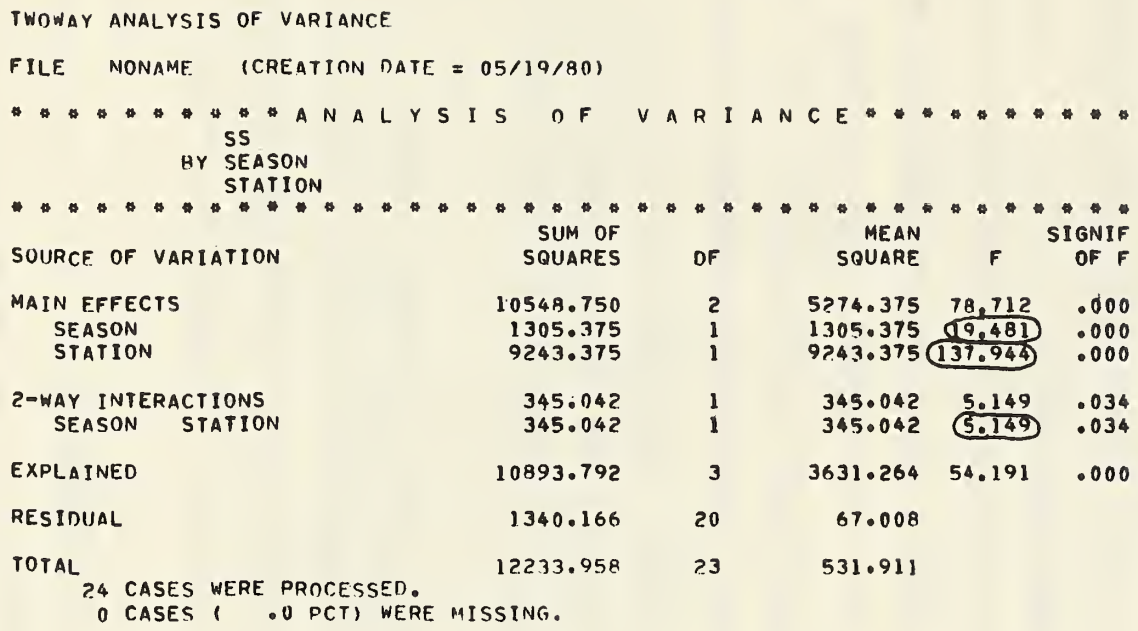
4. Define the critical regions, F_c.

   Since the degrees of freedom for the season, station and interaction are all equal to 1, we find from Table A-4

$$F_c = F_{.01(1,20)} = 8.10$$

5. Reject or accept the null hypothesis.

   Since $F_s > F_c$ for both the seasons and the stations, we do not accept the null hypothesis established in step 1 of the solution. These results indicate that the sediment concentration differs significantly ($\alpha = 0.01$) between stations and between seasons. The interaction effect between season and station is not significant.



```
TWOWAY ANALYSIS OF VARIANCE

FILE   NONAME   (CREATION DATE = 05/19/80)

* * * * * * * * * * A N A L Y S I S   O F   V A R I A N C E * * * * * * * * * *
               SS
            BY SEASON
               STATION
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
                                     SUM OF               MEAN              SIGNIF
SOURCE OF VARIATION                 SQUARES     DF       SQUARE       F      OF F

MAIN EFFECTS                       10548.750     2      5274.375   78.712    .000
   SEASON                           1305.375     1      1305.375   19.481    .000
   STATION                          9243.375     1      9243.375  137.944    .000

2-WAY INTERACTIONS                   345.042     1       345.042    5.149    .034
   SEASON   STATION                  345.042     1       345.042    5.149    .034

EXPLAINED                          10893.792     3      3631.264   54.191    .000

RESIDUAL                            1340.166    20        67.008

TOTAL                              12233.958    23       531.911
     24 CASES WERE PROCESSED.
      0 CASES (   .0 PCT) WERE MISSING.
```

Table 32. Results of the analysis of variance for Example 9f using the ANOVA subprogram from SPSS (1975).

11.0 Regression and Correlation

11.1 Introduction

Regression and correlation are powerful statistical methods which are commonly used in water quality data analysis. With regression, we establish a functional relation of one variable upon another. Correlation, which is often confused with regression, is a measurement of the amount of association between two variables.

This section begins with a detailed discussion of simple linear regression. Initially, the assumptions underlying simple linear regression are examined. Next, the procedures of how to compute the (1) regression line, (2) significance of the regression line, (3) correlation coefficients and (4) confidence limits about the regression line and a point estimate are clearly outlined. This is followed by a brief discussion of covariance in which I review how it can be used to test (1) if the simple linear regression lines are significantly different, and (2) if their slopes are the same.

Next, multiple linear regression is covered. The procedures of how to compute (1) the regression line, (2) the significance of the regression line and (3) the coefficient of multiple determination are covered. Finally, a brief discussion of curvilinear regressions is presented.

11.2 Simple Linear Regression

11.2.1 Assumptions

There are several assumptions underlying the method of simple linear regression. They are listed below along with the methods to test each.

1. The relationship of the two variables, Y to X, is linear, that is $Y = a + bX$. To test this assumption, the first step is to develop a scatter diagram. A scatter diagram is simply a plot of the raw data. By observation, you should be able to tell whether or not a strong linear relationship exists between Y and X. If the relationship does not appear to be linear, then another model should be considered. However, if it does appear to follow a linear relation, then the next step is to develop the regression line and test its significance (these procedures are outlined in Sections 11.2.2 and 11.2.3, respectively).
2. The departures (errors or residuals) of the sample observations from the regression line are independent. To test this assumption, a time-sequence plot of departures, commonly called a residual plot, is developed (Figure 9). If the residuals are independent, then the points should be evenly scattered around the zero residual line.
3. The variance of the observed Y values around the regression line is constant over the range of X values. This assumption can be tested by observation of a plot of the residuals (errors) against the predicted Y values, commonly denoted $\hat{Y}$ (Figure 10). If the variance is homogeneous over the range of X values, the points should be evenly scattered around the zero residual line.
4. The values of X are assumed to be measured without error.
5. Although normality of the data is not required for the regression procedure, the F test for significance of the regression line and the t-test for computation of the confidence limits about the

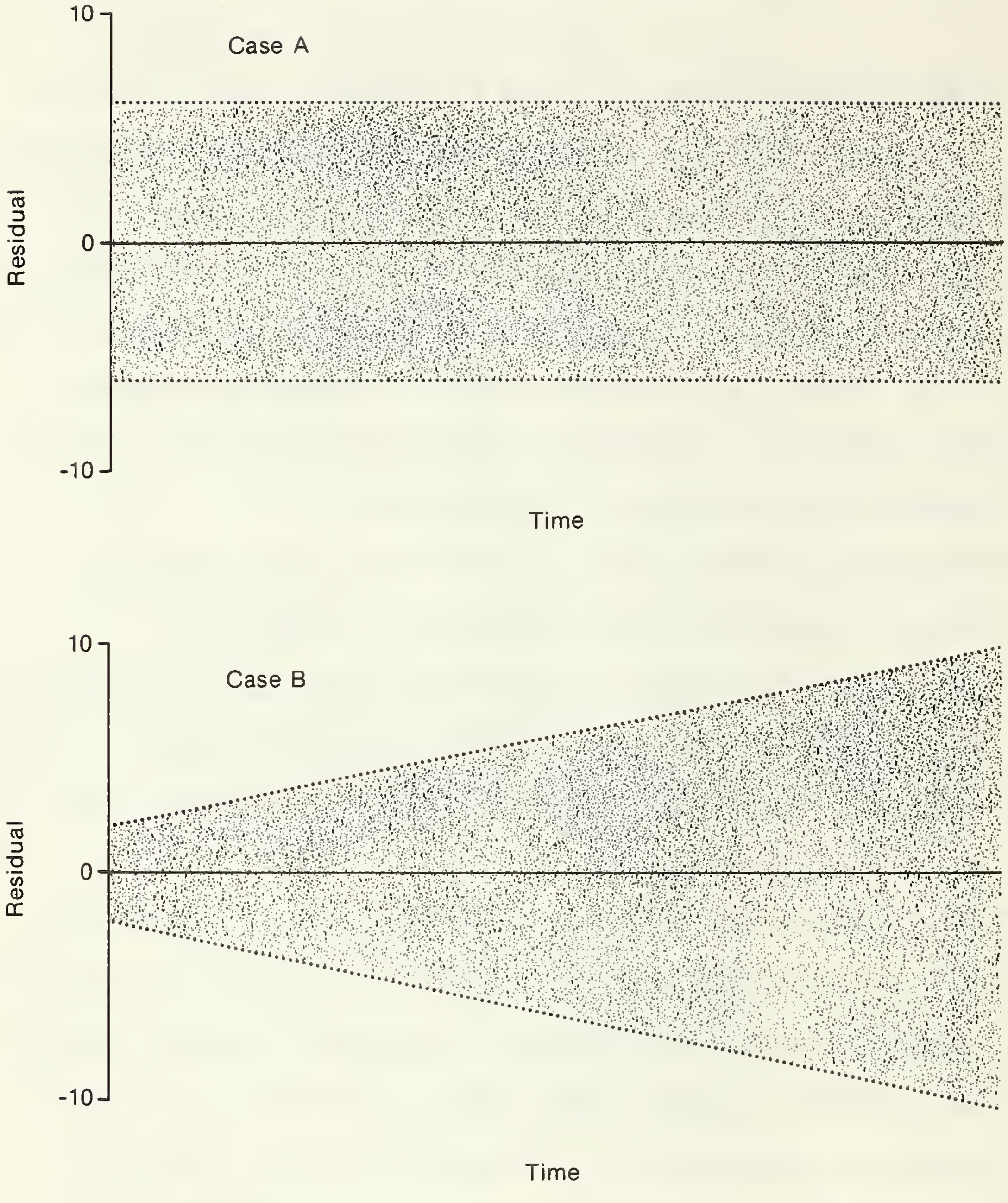


Figure 9. Testing for independence of the residuals. Case A represents independence while case B illustrates non-independence of the residuals.

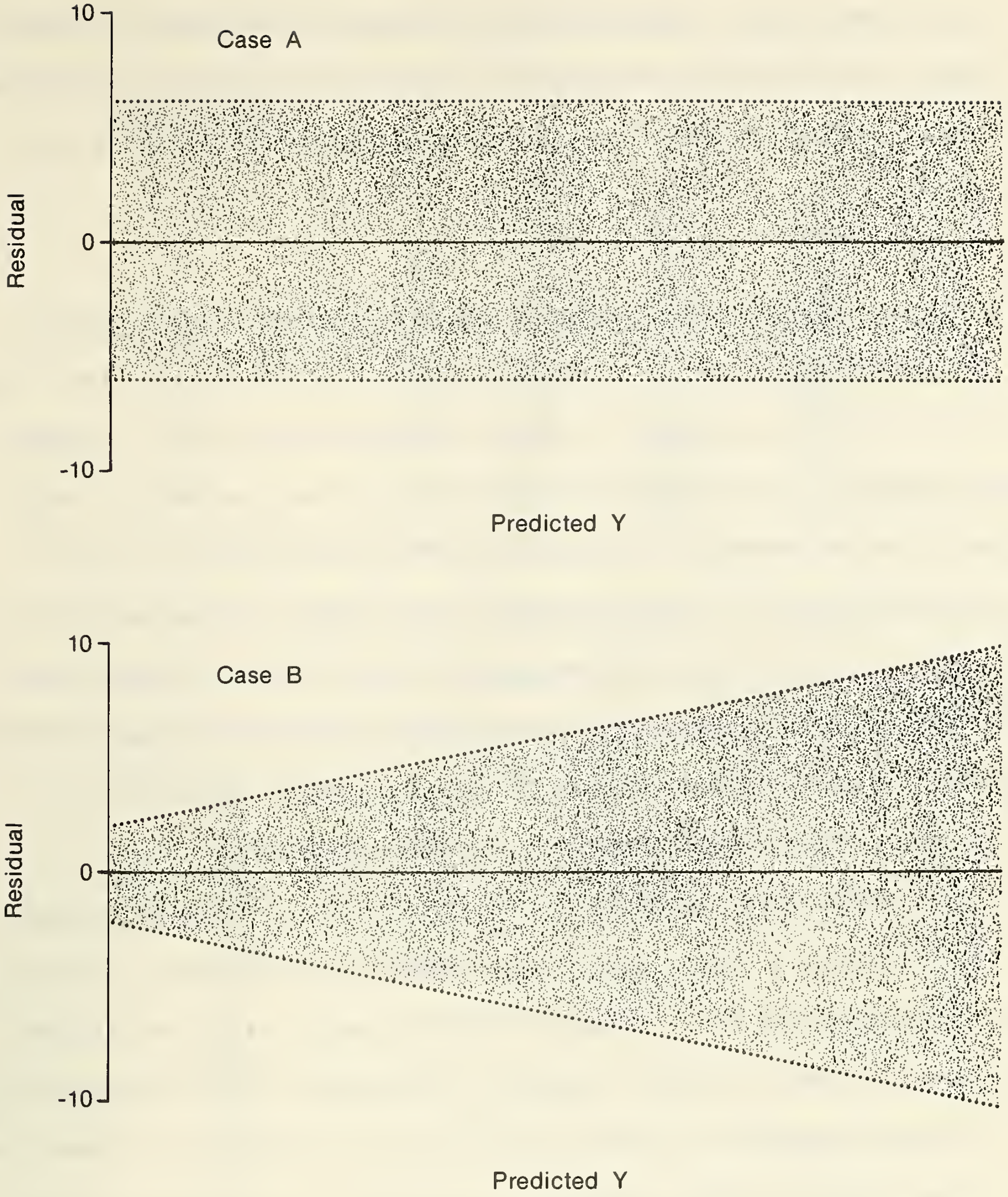


Figure 10. Plots testing for homogeneity in the variances. Plot A represents homogeneity while Plot B is heterogeneous.

regression line do assume that the variation of the observed Y values about the regression line follows a normal distribution. This can be tested using the standard tests for normality discussed in Section 5.1.3.

11.2.2 Least Squares Determination of the Simple Linear Regression Line

The purpose of regression, as you recall, is to establish a functional relationship of one variable with one or more other variables. In simple linear regression, we estimate the relationship of one variable, Y, with another, X, by expressing Y in terms of a linear function of X. For illustrative purposes, consider the data for two water quality variables, X and Y, which are paired by sampling date and were collected at the same station (Table 33). Because of budgetary cutbacks we can no longer afford to sample both parameters as frequently as in the past. We would like to know if we can predict Y accurately if we only measure X.

It is obvious from Figure 11, which is a scatter diagram of the data from Table 33, that a strong linear relationship exists between X and Y. A straight line could be fit through this data by eye to show the relationship between the two variables. However, this method is not very accurate.

A simple mathematical procedure we can use to establish the regression line is the method of least squares. As you recall, the general equation of a straight line is

$$\hat{Y} = a + bX$$

where: a = the value of the Y intercept when X = 0, and

b = a coefficient establishing the slope of the line.

Table 33. Data for two water quality variables, X and Y.[a/]

Sample Number	X	Y	XY	X^2
1	16	18	288	256
2	78	68	5304	6084
3	89	92	8188	7921
4	45	48	2160	2025
5	63	45	2835	3969
6	71	65	4615	5041
7	97	80	7760	9409
8	112	105	11760	12544
9	34	28	952	1156
10	120	108	12960	14400
10	725	657	56822	62805

The objective of the least squares method is to establish the values for the coefficients "a" and "b," which will yield a line for which the sum of the squared deviations from the observed Y values to the straight line

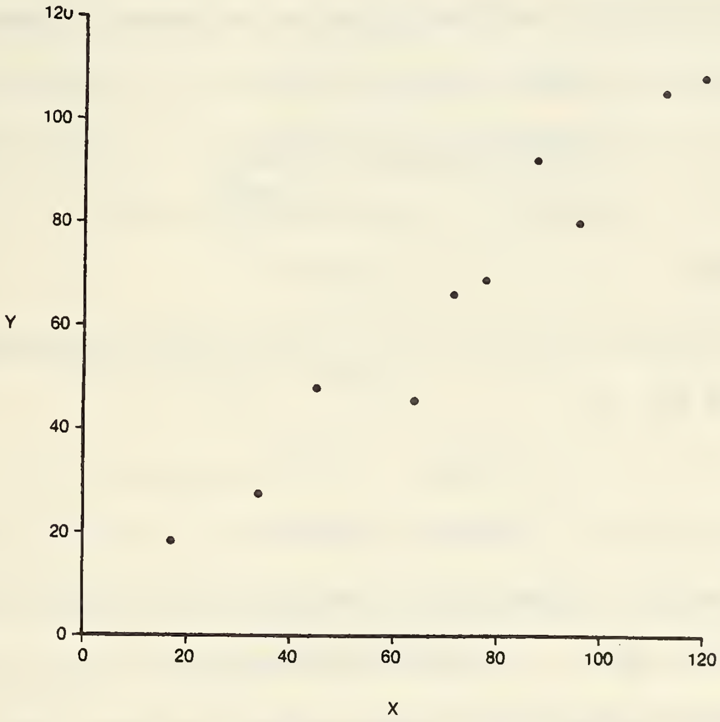


Figure 11. Scatter diagram of X vs. Y for the data from Table 33.

[a/] Also included is other information, XY and X^2, necessary to determine the linear regression line.

is the least possible. The method involves the use of two normal equations (Equations 34 and 35) which are solved simultaneously. The procedure for solving for the coefficients "a" and "b" is as follows.

$$Y = an + bx \qquad (34)$$

$$XY = a \quad \Sigma X + b\Sigma X^2 \qquad (35)$$

1. Substitute the appropriate values from Table 33 into the normal equations.

$$657 = 10a + 725\ b \qquad (36a)$$

$$56822 = 725a + 62805\ b \qquad (36b)$$

2. Divide Equation 36a by 10 and Equation 36b by 725 to bring the value of "a" to 1.0 in each.

$$65.7 = a + 72.5\ b \qquad (36c)$$

$$78.375 = a + 86.628\ b \qquad (36d)$$

3. Subtract Equation 36d from Equation 36c and solve for b.

$$65.7 = a + 72.5\ b \qquad (36c)$$

$$\underline{-78.375 = a + 86.628\ b} \qquad (36d)$$

$$-12.675 = \quad -\ 14.128\ b \qquad (37)$$

Therefore: $b = 0.897$

4. Now, substitute the value of "b" into either Equation 36a or 36b and solve for "a".

$$a = 0.656$$

$$\hat{Y} = 0.656 + 0.897\ X \qquad (38)$$

Using Equation 38 we can now predict a Y value for any value of X. A <u>note of caution</u> is in order here. You should be very careful about extrapolating your estimates beyond the region in which the linear relation was developed.

11.2.3 Testing the Significance of the Simple Linear Regression Line

At this point in our regression analysis we should ask "How well does the regression line fit the data?" To answer this question we need to consider the total variation in the Y data about its mean. By fitting the regression line to the data we have, in effect, attempted to explain part of this variation by the linear association of Y with X. The portion of the variation that remains, that of Y about the regression line, is called the residual or unexplained variation. When we test significance of the regression line, we are seeking to find if the portion of the variation of Y that is explained by the regression line is significantly greater than the portion of the variation of Y that is unexplained.

To test the significance of the regression line we use anova. The total sum of squares for Y, corrected for the mean, is denoted by SS_{total} and estimates the amount of variation of individual Y_i's about Y. The amount of variation in Y that is associated with the regression on X is called the reduction or regression sum of squares, SS_{reg} (Equation 39).

$$SS_{reg} = \frac{(\Sigma xy)^2}{\Sigma x^2} \qquad (39)$$

where

$$\Sigma xy = \Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{n}$$

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

The portion of the total variation in Y that is not associated with the regression is called the residual sum of squares, SS_{res} (Equation 40).

$$SS_{res} = SS_{total} - SS_{reg} \qquad (40)$$

The anova table for the significance of regression line test is given in Table 34. As in the anova testing discussed earlier, we use the

Table 34. The anova table for the significance of regression line test.

Source of Variation	Degrees of Freedom	SS	MS	F_s
Regression	1	SS_{reg}	$s^2_{\hat{Y}}$	$s^2_{\hat{Y}}/s^2_{YX}$
Residual (error)	n-2	SS_{res}	s^2_{YX}	
Total	n-1	SS_{tot}	S^2_{YX}	

residual or error variation as the standard for testing the variation explained by the regression. The calculated F statistic, F_S, is compared with the tabular F, $F_{\alpha(\nu_1, \nu_2)}$, to test for significance.

Freese (1962) points out that if there is a significant difference, this does not mean that the line we fitted gives the best possible description of the data nor does it mean we have found the true mathematical relationship between the two variables. All it allows us to do is state, with a particular degree of probability $(1 - \alpha)$, that the part of the variation in Y that is explained by the fitted line is signiticantly greater than the part that is unexplained.

11.2.4 Correlation

The term used to indicate the degree of correlation is the coefficient of correlation, r. The coefficient of correlation varies between -1.0 and 1.0 and measures the amount of association between two variables. If all the observed Y values lie on the regression line, then $r = \pm 1.0$ depending on the slope of the line. If r = 0, there is no correlation at all between the two variables. In other words, when r = 0, a straight line equal to $\bar{Y}$ or $\bar{X}$ would describe the relationship equally well. Figure 12 illustrates several different examples of coefficients of correlation.

Closely associated with the coefficient of correlation is the coefficient of determination, denoted by r^2. It is a measure of the proportion of the total variation in Y that is associated with the regression of Y on X. Consequently, a $r^2 = 0.64$ means that 64 percent of the variance in variable Y was associated with X. The coefficient of determination can be found using Equation 41.

$$r^2 = \frac{SS_{reg}}{SS_{tot}} \tag{41}$$

11.2.5 Setting Confidence Intervals About a Simple Linear Regression Line and a Point Estimate.

Confidence limits on the regression line can be established by specifying several values over the range of X and computing the lower (L) and upper (U) limit using Equations 42 and 43.

$$L = \hat{Y}_i - t_{\alpha(\nu)} \sqrt{s^2_{YX}\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma x^2}\right)} \tag{42}$$

$$U = \hat{Y}_i + t_{\alpha(\nu)} \sqrt{s^2_{YX}\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma x^2}\right)} \tag{43}$$

where Y_i is the predicted value of Y for a particular X_i, ν is the degrees of freedom for the residual mean square (s^2_{YX}) and the other terms are as previously defined.

It should be pointed out that these are confidence limits on the regression of Y on X. In other words, they indicate the limits for a band which will cover the true mean of Y for a given X, unless the α in-one

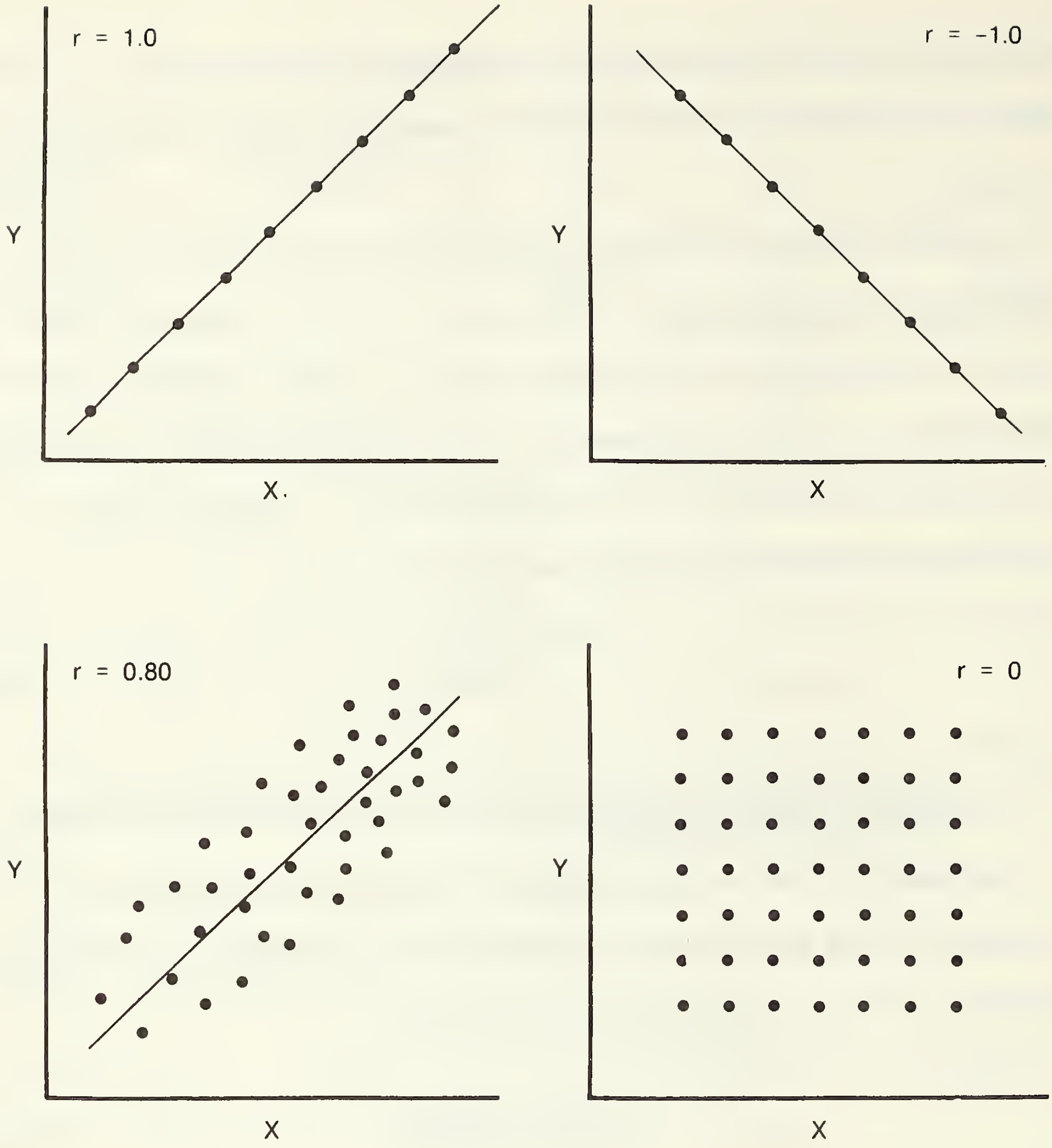


Figure 12. Several different examples of coefficients of correlation.

chance has occurred (Freese, 1962). These limits do not apply to a single predicted value of Y. The limits which will cover a single Y are given by Equations 44 and 45.

$$L = \hat{Y} - t_{\alpha(\nu)} \sqrt{s^2_{YX}\left(1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma x^2}\right)} \tag{44}$$

$$U = \hat{Y} + t_{\alpha(\nu)} \sqrt{s^2_{YX}\left(1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma x^2}\right)} \tag{45}$$

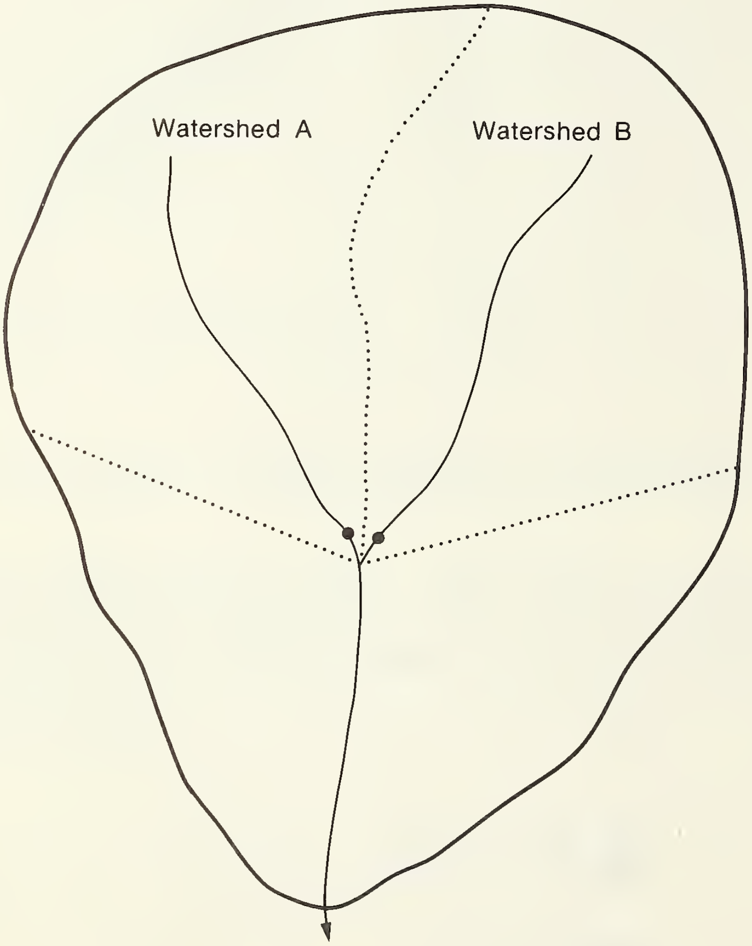
11.2.6 Example

An example of how to apply simple linear regression analysis is presented in Example 10.

EXAMPLE 10
Simple Linear Regression Analysis

Problem:

Suspended solids concentrations were monitored at Stations A and B (see illustration below) over a two-year period. The results are tabulated below the illustration. You are to (1) develop a scatter diagram using the raw data (data from Watershed B are to be assigned X values), (2) fit a simple linear regression line to the data, (3) test the significance of the regression (P = 0.05), (4) determine the coefficient of determination and (5) set the 95% confidence limits about the predicted values.



Suspended Solids Concentrations (mg/l)

Subwatershed A	Subwatershed B
14	41
92	260
54	230
66	170
109	351
62	249
28	61
36	87
41	180
23	104
5	20
97	289

Solution:

1. The scatter diagram was constructed by hand and is illustrated in Figure 14. It is readily apparent that a linear regression can be fit to these data.

2. Parts (2) through (5) of the Example were solved using the GLM program from SAS (1979). [It should be noted that the REGRESSION program from SPSS (1975) could also have been used to solve this problem. However, REGRESSION will not calculate the confidence limits about the predicted values like GLM does.] The results of the GLM program are presented in Table 35.

The simple linear regression line for this example problem is

$$SS_A = 1.5844 + 0.2977\ SS_B.$$

The output clearly indicates that the regression is highly significant. $F_s = 84.08$ while $F_c = F_{.05(1,10)} = 4.96$.

The coefficient of determination, r^2, is equal to 0.89.

The confidence limits about the predicted values are tabulated at the bottom of Table 35 and have been plotted in Figure 13. The plotted confidence limits are not straight lines because the greater the distance from the mean of X and Y the wider the range at a given probability level.

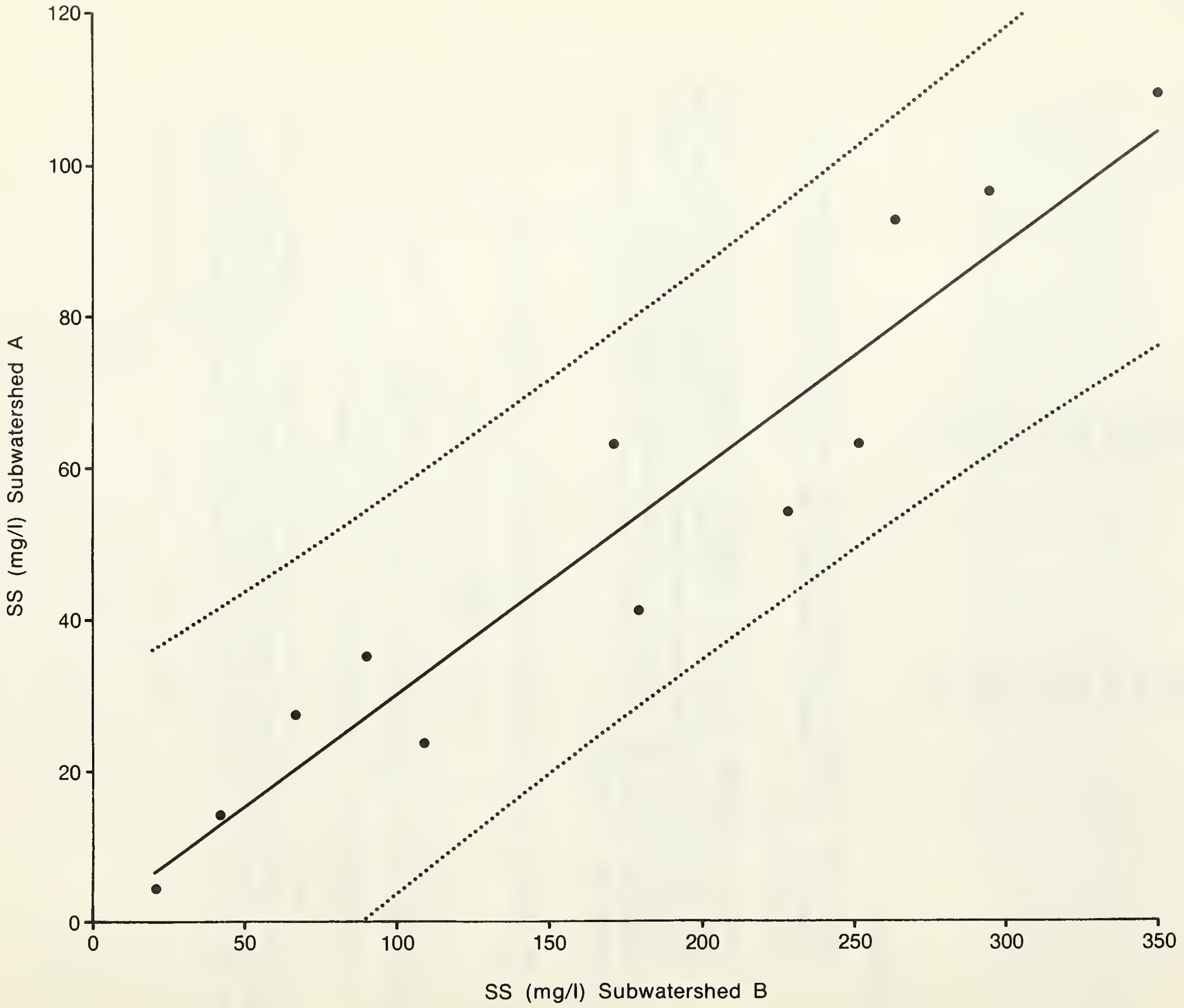


Figure 13. The scatter diagram of the data from Example 10, the simple linear regression line through the data and the 95% confidence limits about the regression line.

```
                                        GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: SWA

SOURCE                 DF        SUM OF SQUARES         MEAN SQUARE       F VALUE         PR > F        R-SQUARE          C.V.

MODEL                   1        11278.87920991      11278.87920991         84.08         0.0001        0.893713       22.1660

ERROR                  10         1341.37079009        134.13707901                       STD DEV                     SWA MEAN

CORRECTED TOTAL        11        12620.25000000                                        11.58175630                 52.25000000


SOURCE                 DF            TYPE I SS      F VALUE      PR > F        DF            TYPE IV SS      F VALUE      PR > F

SWB                     1       11278.87920991        84.08      0.0001         1       11278.87920991        84.08      0.0001


                                     T FOR H0:      PR > |T|        STD ERROR OF
PARAMETER              ESTIMATE     PARAMETER=0                       ESTIMATE

INTERCEPT            1.58439070            0.25       0.8112          6.45808557
SWB                  0.29774109            9.17       0.0001          0.03246985


OBSERVATION           SWB       OBSERVED          PREDICTED           RESIDUAL       LOWER 95% CL       UPPER 95% CL
                                   VALUE              VALUE                            INDIVIDUAL         INDIVIDUAL

        1              41      14.00000000       13.79177550         0.20822450      -14.64707095        42.23062196
        2             260      92.00000000       78.99707484        13.00292516       51.36230293       106.63184675
        3             230      54.00000000       70.06484206       -16.06484206       42.85860976        97.27107436
        4             170      66.00000000       52.20037648        13.79962352       25.34072920        79.06002377
        5             351     109.00000000      106.09151430         2.90848570       76.21506255       135.96796605
        6             249      62.00000000       75.72192282       -13.72192282       48.26341887       103.18042677
        7              61      28.00000000       19.74659736         8.25340264       -8.25015812        47.74335284
        8              87      36.00000000       27.48786578         8.51213422       -0.03746520        55.01319675
        9             180      41.00000000       55.17778741       -14.17778741       28.30872296        82.04685187
       10             104      23.00000000       32.54946435        -9.54946435        5.26657659        59.83235212
       11              20       5.00000000        7.53921255        -2.53921255      -21.43442693        36.51285203
       12             289      97.00000000       87.63156654         9.36843346       59.42953671       115.83359637

                      SUM OF RESIDUALS                               -0.00000000
                      SUM OF SQUARED RESIDUALS                     1341.37079009
                      SUM OF SQUARED RESIDUALS - ERROR SS            -0.00000000
                      PRESS STATISTIC                              1757.08534267
                      FIRST ORDER AUTOCORRELATION                    -0.33956081
                      DURBIN-WATSON D                                 2.61365806
```

Table 35. The GLM output from SAS (1979) for Example 10.

11.2.7 Analysis of Covariance to Compare the Simple Linear Regression Lines Developed from Several Groups of Data.

In some cases, we may want to compare two or more simple linear regression lines and determine if they differ in either their slope or in their level. The statistical method to use in this situation is the analysis of covariance. The procedures necessary for this analysis can best be described using an illustrative example. Much of the discussion that follows has been taken directly from Freese (1967).

Consider two groups, A and B, composed of X and Y data. This could represent a control watershed and a treated watershed (X and Y, respectively) where data were collected prior to and following treatment (groups A and B, respectively). Linear regressions of Y on X were fitted for each of the two groups. The basic data and the fitted regressions were as follows:

Group A

										Sum	Mean
Y	3	7	9	6	8	13	10	12	14	82	9.111
X	1	4	7	7	2	9	10	6	12	58	6.444

$n = 9$, $\Sigma Y^2 = 848$, $\Sigma XY = 609$, $\Sigma X^2 = 480$, $\Sigma y^2 = 100.8889$,

$\Sigma xy = 80.5556$, $\Sigma x^2 = 106.2222$, $Y = 4.224 + 0.7584X$

Residual SS = 39.7980, with 7 df.

Group B

														Sum	Mean
Y	4	6	12	2	8	7	0	5	9	2	11	3	10	79	6.077
X	4	9	14	6	9	12	2	7	5	5	11	2	13	99	7.615

$n = 13$, $\Sigma Y^2 = 653$, $\Sigma XY = 753$, $\Sigma X^2 = 951$, $\Sigma y^2 = 172.9231$,

$\Sigma xy = 151.3846$, $\Sigma x^2 = 197.0769$, $Y = 0.228 + 0.7681X$

Residual SS = 56.6370 with 11 df.

In testing for common regressions the procedure is to test first for common slopes. If the slopes differ significantly, the regressions are different and no further testing is needed. If the slopes are not significantly different, the difference in level is tested. The analysis table is as follows:

Table 36. Analysis of Covariance Table (after Freeze, 1967).

						Residuals		
Line	Group	df	Σy^2	Σxy	Σx^2	df	SS	MS
1	A	8	100.8889	80.5556	106.2222	7	39.7980	
2	B	12	172.9231	151.3846	197.0769	11	56.6370	
3					Pooled residuals	18	96.4350	5.3575
4	Difference for testing common slopes					1	0.0067	0.0067
5	Common slope	20	273.8120	231.9402	303.2991	19	96.4417	5.0759
6	Difference for testing levels					1	80.1954	80.1954
7	Single regression	21	322.7727	213.0455	310.5909	20	176.6371	

The first two lines in this table contain the basic data for the two groups. To the left are the total df for the groups (8 for A and 12 for B). In the center are the corrected sums of squares and products. The right side of the table gives the residual sums of squares and df. Since only simple linear regressions have been fitted, the residual df for each group is one less than the total df. The residual sum of squares is obtained by first computing the reduction sum of squares (SS due to regression) for each group (Equation 46). This reduction is then subtracted from the total sum of squares (Σy^2) to give the residuals.

$$SS_{reg} = \frac{(\Sigma xy)^2}{\Sigma x^2} \qquad (46)$$

Line 3 is obtained by pooling the residual df and residual sums of squares for the groups. Dividing the pooled sum of squares by the pooled df gives the pooled mean square.

The left side and center of line 5 are obtained by pooling the total df and the corrected sums of squares and products for the groups. These are the values that are obtained under the assumption of no difference in the slopes of the group regressions. If the assumption is wrong, the residuals about this common slope regression will be considerably larger than the mean square residual about the separate regressions. The residual df and sum of squares are obtained by fitting a straight line to this pooled data. The residual df is one less than the total df. The residual sum of squares is

$$SS_{res} = 273.8120 = \frac{(231.9402)^2}{303.2991} = 96.4417$$

Now the difference between these residuals (line 4 = line 5 - line 3) provides a test of the hypothesis of common slopes. The error term for this test is the pooled mean square from line 3.

$$\text{Test of common slopes: } F_{(1,18)} = \frac{0.0067}{5.3575}$$

The difference is not significant.

If the slopes differed significantly, the groups would have different regressions, and we would stop here. Since the slopes did not differ, we now go on to test for a difference in the levels of the regression.

Line 7 is what we would have if we ignored the groups entirely, lumped all the original observations together and fitted a single linear regression. The combined data are as follows:

$$n = (9 + 13) = 22 \text{ (so the df for total = 21)}$$

$$\Sigma Y = (82 + 79) = 161, \ \Sigma Y^2 = (848 + 653) = 1{,}501$$

$$\Sigma y^2 = 1{,}501 - \frac{(161)^2}{22} = 322.7727$$

$$\Sigma X = (58 + 99) = 157, \ \Sigma X^2 = (480 + 951) = 1{,}431$$

$$\Sigma x^2 = 1{,}431 - \frac{(157)^2}{22} = 310.5909$$

$$\Sigma XY = (609 + 753) = 1{,}362, \ \Sigma xy = 1{,}362 - \frac{(157)(161)}{22} = 213.0455$$

From this we obtain the residual values on the right side of line 7.

$$SS_{res} = 322.7727 - \frac{(213.0455)^2}{310.5909} = 176.6371$$

If there is a real difference among the levels of the groups, the residuals about this single regression will be considerably larger than the mean square residual about the regression that assumed the same slopes but different levels. This difference (line 6 = line 7 - line 5) is tested against the residual mean square from line 5.

$$\text{Test of levels:} \quad F_{s(1,19)} = \frac{80.1954}{5.0759} = 15.80^{**}$$

As the levels differ significantly, the groups do not have the same regressions.

11.3 Multiple Regression

11.3.1 Assumptions

The assumptions underlying the methods of fitting a multiple regression are the same as those for a simple linear regression (Section 11.2.1).

11.3.2 Least Squares Determination of the Multiple Regression Line

In some situations, the dependent variable (Y) is related to more than one independent variable (X_1, X_2, X_3, . . . , X_n). As we have shown, we could fit a simple linear regression to the data using only one independent variable. However, the amount of variation explained by a regression would probably be much better if we used all the independent variables together. The statistical method which enables us to do this is called multiple regression.

As was the case with simple linear regression, the best regression line can be determined using the method of least squares. Again, a series of normal equations are established and solved simultaneously. For the general linear model with a constant term

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \qquad (47)$$

it is very easy to develop the normal equations, once you recognize the pattern. Each term in the first row contains an x_1; each term in the second row contains an x_2; and so on through the nth row. Each term in the first column has an x_1 and b_1; each term in the second column has a x_2 and b_2; and so on through the nth column. Each ith row will contain a term of the form $(\Sigma x_i^2)b_i$. On the right side of the equations, every equation has a term of $\Sigma x_i y$. For the general model presented above (Equation 47) the normal equations would be as follows:

$$(\Sigma x_1^2)b_1 + (\Sigma x_1x_2)b_2 + (\Sigma x_1x_3)b_3 + \ldots + (\Sigma x_1x_n)b_n = \Sigma x_n y \qquad (48)$$

$$(\Sigma x_1x_2)b_1 + (\Sigma x_2^2)b_2 + (\Sigma x_2x_3)b_3 + \ldots + (\Sigma x_2x_n)b_n = \Sigma x_2 y \qquad (49)$$

$$(\Sigma x_1x_3)b_1 + (\Sigma x_2x_3)b_2 + (\Sigma x_3^2)b_3 + \ldots + (\Sigma x_3x_n)b_n = \Sigma x_3 y \qquad (50)$$

$$\vdots$$

$$(\Sigma x_1x_n)b_1 + (\Sigma x_2x_n)b_2 + (\Sigma x_3x_n)b_3 + \ldots + (\Sigma xn^2)b_n = \Sigma x_n y \qquad (51)$$

Once we have obtained the "b" coefficients, the constant term can be found using the general equation

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3 - \ldots - b_n\bar{X}_n. \qquad (52)$$

The mechanics of the least squares determination of the multiple regression line are outlined using an illustrated example. (The example presented here has been taken directly from Freese, 1967. It is included with only minor modification because I feel it presents the concept in a very clear and concise manner.) Consider the data presented in Table

37. With this data we would like to fit an equation of the form

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

Table 37. Data for the example illustrating the least squares determination of the multiple regression line (after Freese, 1967).

	Y	X_1	X_2	X_3
	65	41	79	75
	78	90	48	83
	85	53	67	74
	50	42	52	61
	55	57	52	59
	59	32	82	73
	82	71	80	72
	66	60	65	66
	113	98	96	99
	86	80	81	90
	104	101	78	86
	92	100	59	88
	96	84	84	93
	65	72	48	70
	81	55	93	85
	77	77	68	71
	83	98	51	84
	97	95	82	81
	90	90	70	78
	87	93	61	89
	74	45	96	81
	70	50	80	77
	75	60	76	70
	75	68	74	76
	93	75	96	85
	76	82	58	80
	71	72	58	68
	61	46	69	65
Sums	2,206	1,987	2,003	2,179
Means (n = 28)	78.7857	70.9643	71.5357	77.8214

According to the principle of least squares, the best estimates of the "b" coefficients can be obtained by solving the set of least squares normal equations.

$$b_1 \text{ equation: } (\Sigma x_1^2)b_1 + (\Sigma x_1x_2)b_2 + (\Sigma x_1x_3)b_3 = \Sigma x_1y$$

$$b_2 \text{ equation: } (\Sigma x_1x_2)b_1 + (\Sigma x_2^2)b_2 + (\Sigma x_2x_3)b_3 = \Sigma x_2y$$

$$b_3 \text{ equation: } (\Sigma x_1x_3)b_1 + (\Sigma x_2x_3)b_2 + (\Sigma x_3^2)b_3 = \Sigma x_3y$$

where: $\Sigma x_ix_j = \Sigma X_iX_j - \frac{(\Sigma X_i)(\Sigma X_j)}{n}$

The corrected sums of squares and products are computed in the familiar manner:

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = (65^2 + \ldots + 61^2) - \frac{(2206)^2}{28} = 5{,}974.7143$$

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n} = (41^2 + \ldots + 46^2) - \frac{(1987)^2}{28} = 11{,}436.9643$$

$$\Sigma x_1y = \Sigma X_1Y - \frac{(\Sigma X_1)\,(\Sigma Y)}{n} = (41)(65) + \ldots + (46)(61) - \frac{(1{,}987)(2{,}206)}{28} = 6{,}428.7858$$

Similarly,

$$\begin{aligned}
\Sigma x_1x_2 &= -1{,}171.4642 \\
\Sigma x_1x_3 &= 3{,}458.8215 \\
\Sigma x_2^2 &= 5{,}998.9643 \\
\Sigma x_2x_3 &= 1{,}789.6786 \\
\Sigma x_2y &= 2{,}632.2143 \\
\Sigma x_3^2 &= 2{,}606.1072 \\
\Sigma x_3y &= 3{,}327.9286
\end{aligned}$$

Putting these values in the normal equations gives:

$$11{,}436.9643b_1 - 1{,}171.4642b_2 + 3{,}458.8215b_3 = 6{,}428.7858$$

$$-\ 1{,}171.4642b_1 + 5{,}998.9643b_2 + 1{,}789.6786b_3 = 2{,}632.2143$$

$$3{,}458.8215b_1 + 1{,}789.6786b_2 + 2{,}606.1072b_3 = 3{,}327.9286$$

These equations can be solved by any of the standard procedures for simultaneous equations. One approach is as follows:

1. Divide through each equation by the numerical coefficient of b_1.

$$b_1 - 0.102{,}427{,}897b_2 + 0.302{,}424{,}788b_3 = -0.562{,}105{,}960$$

$$b_1 - 5.120{,}911{,}334b_2 - 1{,}527{,}727{,}949b_3 = -2.246{,}943{,}867$$

$$b_1 + 0.517{,}424{,}389b_2 + 0.753{,}466{,}809b_3 = 0.962{,}156{,}792$$

2. Subtract the second equation from the first and the third from the first so as to leave two equations in b_2 and b_3.

$$5.018{,}483{,}437b_2 + 1.830{,}152{,}737b_3 = 2.809{,}049{,}827$$

$$-0.619{,}852{,}286b_2 - 0.451{,}042{,}021b_3 = -0.400{,}050{,}832$$

3. Divide through each equation by the numerical coefficient of b_2.

$$b_2 + 0.364{,}682{,}430b_3 = 0.559{,}740{,}779$$

$$b_2 + 0.727{,}660{,}494b_3 = 0.645{,}397{,}042$$

4. Subtract the second of these equations from the first, leaving one equation in b_3.

$$-0.362{,}978{,}064b_3 = -0.085{,}656{,}263$$

5. Solve for b_3.

$$b_3 = \frac{-0.085{,}656{,}263}{-0.362{,}978{,}064} = 0.235{,}981{,}927$$

6. Substitute this value of b_3 in one of the equations (say the first) of step 3 and solve for b_2.

$$b_2 + (0.364{,}682{,}430)(0.235{,}981{,}927) = 0.559{,}740{,}779$$

$$b = 0.473{,}682{,}316$$

7. Substitute the solutions for b_2 and b_3 in one of the equations (say the first) of step 1, and solve for b_1.

$$b_1 - (0.102{,}427{,}897)(0.473{,}682{,}316) + (0.302{,}424{,}788)(0.235{,}981{,}927)$$
$$= 0.562{,}105{,}960$$

b_1 = 0.539,257,459

8. As a check, add up the original normal equations and substitute the solutions for b_1, b_2 and b_3.

$13{,}724.3216b_1 = 6{,}617{,}1787b_2 + 7{,}854.6073b_3 = 12{,}388.9287$

$12{,}388.92869 = 12{,}388.9287$, check.

Given the values of b_1, b_2, and b_3 we can now compute

$$a = \hat{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3 = -11.7320$$

Thus, after rounding of the coefficients, the regression equation is

$$Y = -11.732 + 0.539\ X_1 + 0.474\ X_2 + 0.236\ X_3$$

It should be noted that in solving the normal equations more digits have been carried than would be justified by the rules for number of significant digits. Unless this is done, the rounding errors may make it difficult to check the computations.

11.3.3 Testing the Significance of the Multiple Regression Line

At this point in our regression analysis we should ask, "How well does the regression line fit the data?" The analysis of variance procedure to be used here is similar to that outlined for the significance test of the simple linear regression line. However, in this case the degrees of freedom for the reduction are equal to the number of independent variables fitted. The reduction sum of squares for any least squares regression can be found using the general equation:

SS_{reg} = (est. coefficients)(right side of their respective normal equations) (53)

Therefore, for Freese's example

$$SS_{reg} = b_1(\Sigma x_1 y) + b_2(\Sigma x_2 y) + b_3(\Sigma x_2 y)$$

The anova table for the test of significance is as follows in Table 38:

Table 38. ANOVA results for the test of significance of the multiple regression developed from the data given in Table 36 (after Freese, 1967).

Source	df	SS	MS
Reduction due to X_1, X_2, and X_3 . . .	3	5,498.9389	1,832.9796
Residuals	24	475.7754	19.8240
Total	27	5,974.7143	

To test the significance of the regression we compute F_s where

$$F_s = \frac{\text{Reduction MS}}{\text{Residual MS}}$$

For the case at hand $F_s = 92.46$, which is significant at the 0.01 level.

In some instances we would like to know the contribution each independent variable makes in the prediction of the dependent variable. In other words, what portion of the total SS can be attributed to each individual independent variable. The statistical method to use for this type of analysis is stepwise regression. The procedures for stepwise regression are not outlined here, but can be found in any good statistics text.

11.3.4 Coefficient of Multiple Determination

The coefficient of multiple determination is calculated in the same manner as that for the simple linear regression.

$$r^2 = \frac{SS_{reg}}{SS_{tot}} \tag{54}$$

For our illustrative example, $r^2 = 0.92$, which means 92 percent of the variation in Y is associated with the regression.

11.4 Curvilinear Regression

Only a very brief discussion of curvilinear regression is presented here. In general, many of the curvilinear relationships can be handled using the regression methods already presented. Consider the simple power function

$$Y = aX^b \tag{55}$$

Water quality data collected from streams will very often follow this relationship. This function can be linearized using a simple log transformation (Equation 56). In addition, many other curvilinear relationships can be linearized. Chow (1964) presents a very extensive tabulation of transformations for linearization of different equations (Table 39).

$$\log Y = \log a + b \log X \tag{56}$$

However, some curvilinear functions, such as

$$Y = a + b^X$$

or

$$Y = a(X-b)^2$$

cannot be fitted by the methods already described. To regress these functions requires procedures beyond the scope of this paper.

Table 39. Transformations of linearization of different functions (after Chow, 1967).

	Type of function	Straight-line coordinate		Equation in linear form
		Abscissa	Ordinate	
1	$y = a + bx$	x	y	$[y] = a + b[x]$
2	$y = be^{ax}$	x	$\log y$	$[\log y] = \log b + (a \log c)[x]$
3	$y = ax^b$	$\log x$	$\log y$	$[\log y] = \log a + b[\log x]$
4	$y = a_0 + a_1x + a_2x^2$	$x - x_0$	$\dfrac{y - y_0}{x - x_0}$	$\left[\dfrac{y - y_0}{x - x_0}\right] = a_1 + 2a_1x_0 + a_2[(x - x_0)]$
5	$y = a + b/x$	$1/x$	y	$[y] = a + b[1/x]$
6	$y = x/(a + bx)$	x	x/y	$[x/y] = a + b[x]$
7	$y = a/(b + cx)$	x	$1/y$	$[1/y] = (b/a) + (c/a)[x]$
8	$y = c + be^{ax}$	x	$\log \dfrac{\Delta y}{\Delta x}$	$\left[\log \dfrac{dy}{dx}\right] = \log (ab) + (a \log e)[x]$
9	$y = c + ax^b$	$\log x$	$\log \dfrac{\Delta y}{\Delta x}$	$\left[\log \dfrac{dy}{dx}\right] = \log (ab) + (b - 1)[\log x]$
10	$y = c + \dfrac{b}{x - a}$	$x - x_0$	$\dfrac{x - x_0}{y - y_0}$	$\left[\dfrac{x - x_0}{y - y_0}\right] = -\dfrac{a - x_0}{c - y_0} + \dfrac{1}{c - y_0}[x - x_0]$
11	$y = c + \dfrac{x}{a + bx}$	x	$\dfrac{x - x_0}{y - y_0}$	$\left[\dfrac{x - x_0}{y - y_0}\right] = (a + bx_0) + \dfrac{b(a + bx_0)}{a}[x]$
12	$y = d + cx + be^{ax}$	x	$\log \dfrac{\Delta^2 y}{\Delta x^2}$	$\left[\log \dfrac{d^2y}{dx^2}\right] = \log (a^2b) + (a \log c)[x]$ $[y - be^{ax}] = d + c[x]$
13	$y = dc^xb^m$, where $m = a^x$	x	$\log \dfrac{\Delta^2(\log y)}{\Delta x^2}$	$\left[\dfrac{\log d^2(\log y)}{dx^2}\right] = \log \left[\dfrac{(\log b)(\log a)^2}{(\log c)^2}\right] + (\log a)[x]$
14	$y = de^{cx} + be^{ax}$	$\dfrac{y_{k+1}}{y_k}$	$\dfrac{y_{k+2}}{y_k}$	$[\log y - a^x \log b] = \log d + (\log e)[x]$ $\left[\dfrac{y_{k+2}}{y_k}\right] = -e^{(a+c)\Delta x} + (e^{a\Delta x} + e^{c\Delta x})\left[\dfrac{y_{k+1}}{y_k}\right]$ $[ye^{-cx}] = d + b[e^{(a-c)x}]$
15	$y = e^{ax}(d \cos bx + c \sin bx)$	$\dfrac{y_{k+1}}{y_k}$	$\dfrac{y_{k+2}}{y_k}$	$\left[\dfrac{y_{k+2}}{y_k}\right] = -e^{2a\Delta x} = (2e^{a\Delta x} \cos b\,\Delta x)\left[\dfrac{y_{k+1}}{y_k}\right].$ $\left[\dfrac{yc^{-ax}}{\cos bx}\right] = d + c[\tan bx]$

REMARK: In types 14 and 15, y_k, y_{k+1}, and y_{k+2} are consecutive values for an increment Δx.

12.0 Literature Cited

Averett, R. 1979. The use of select parametric statistical methods for the analysis of water quality data. Presented at USGS - BLM Conference on Water-Quality in Energy Areas. January 10-11, Denver, Colorado. 16p.

Calquhoun, D. 1971. Lectures on biostatistics. Claredon Press.

Elliott, J.M. 1971. Some methods for the statistical analysis of Benthic Invertebrates. Fresh Water Biol. Assoc. Sci. Pub. 25. 144p.

Freese, F. 1967. Elementary statistical methods for foresters. Agriculture Handbook 317. USDA - Forest Service. 87p.

Glass, G.V., P.D. Peckham and J.R. Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Rev. Educ. Res. 42:237-288.

Ingwersen, J.B. 1981a. Statistical analysis using SAS at the USEPA National Computer Center. WSDG Application Document WSDG-AD-00001, USDA Forest Service, 51p.

Ingwersen, J.B. 1981b. Statistical analysis using SPSS at the USDA-Fort Collins Computer Center. WSDG Application Document WSDG-AD-00002, USDA Forest Service, 9p.

Nash, A.J. 1965. Statistical techniques in forestry. Lucas Brothers Publishers. Columbia, Missouri. 146 p. Statistical Analysis System. 1979.

SAS Users Guide. SAS Institute Inc., 1980. P.O. Box 10066, Raleigh, North Carolina 17605. 494 p.

Shapiro, S.S. and M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika. Vol. 52, pp. 591-611.

Sokal, R.R. and F.J. Rohlf. 1969. Biometry. W.H. Freeman and Company. San Francisco. 776 p.

Statistical Package for the Social Sciences. 1975. SPSS: Statistical Package for the Social Sciences. McGraw-Hill, Inc. New York, NY 675p.

Stephans, M.A. 1974. Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables. J. American Statistical Association, 69:730.

Steel, R.G.D. and J. H. Torrie. 1960. Principles and procedures of statistics with special reference to the biological sciences. McGraw-Hill Book Company, Inc. 481 p.

APPENDIX A

STATISTICAL TABLES

Table A-1. Probability of a random value of $z = (X-\mu)/\sigma$ being greater than the values tabulated in the margins (Steel and Torrie, 1960).

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.9	.0000									

Table A-2. Values of t (Steel and Torrie, 1960).

df	Probability of a larger value of *t*, sign ignored								
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291
	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.0005
df	Probability of a larger value of *t*, sign considered								

Source: This table is abridged from Table III of Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Oliver and Boyd Ltd., Edinburgh, 1949, by permission of the authors and publishers.

Table A-3. Values of χ^2 (Steel and Torrie, 1960).

df	Probability of a larger value of χ^2												
	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	$.0^4393$	$.0^3157$	$.0^3982$	$.0^2393$	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0

SOURCE: This table is abridged from "Table of percentage points of the χ^2 distribution," *Biometrika*, 32: 188–189 (1941), by Catherine M. Thompson. It is published here with kind permission of the author and the editor of *Biometrika*.

Table A-4. Values of F (Steel and Torrie, 1960).

Denominator *df*	Probability of a larger *F*	Numerator *df*								
		1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
	.010	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
	.005	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72

Numerator *df*											
10	12	15	20	24	30	40	60	120	∞	*P*	*df*
60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33	.100	1
241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3	.050	
968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018	.025	
6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	.010	
24224	24426	24630	24836	24940	25044	25148	25253	25359	25465	.005	
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	.100	2
19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50	.050	
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50	.025	
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	.010	
199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	.005	
5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13	.100	3
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	.050	
14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90	.025	
27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13	.010	
43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83	.005	
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76	.100	4
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	.050	
8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26	.025	
14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46	.010	
20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32	.005	
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10	.100	5
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36	.050	
6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02	.025	
10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	.010	
13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14	.005	
2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72	.100	6
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67	.050	
5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85	.025	
7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	.010	
10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88	.005	
2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47	.100	7
3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	.050	
4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14	.025	
6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	.010	
8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08	.005	
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29	.100	8
3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93	.050	
4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67	.025	
5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	.010	
7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95	.005	
2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16	.100	9
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71	.050	
3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33	.025	
5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	.010	
6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19	.005	
2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06	.100	10
2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	.050	
3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08	.025	
4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	.010	
5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64	.005	
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97	.100	11
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40	.050	
3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88	.025	
4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	.010	
5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23	.005	
2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90	.100	12
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30	.050	
3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72	.025	
4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	.010	
5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90	.005	
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85	.100	13
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21	.050	
3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60	.025	
4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	.010	
4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65	.005	
2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80	.100	14
2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13	.050	
3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49	.025	
3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00	.010	
4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44	.005	

Denominator *df*	Probability of a larger *F*	Numerator *df*								
		1	2	3	4	5	6	7	8	9
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52

Numerator *df*											
10	12	15	20	24	30	40	60	120	∞	*P*	*df*
2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76	.100	15
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07	.050	
3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40	.025	
3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	.010	
4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26	.005	
2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72	.100	16
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01	.050	
2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32	.025	
3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	.010	
4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11	.005	
2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69	.100	17
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96	.050	
2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25	.025	
3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	.010	
4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98	.005	
1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66	.100	18
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92	.050	
2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19	.025	
3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	.010	
4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87	.005	
1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63	.100	19
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88	.050	
2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13	.025	
3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	.010	
3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78	.005	
1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61	.100	20
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84	.050	
2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09	.025	
3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	010	
3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69	.005	
1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59	.100	21
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	.050	
2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04	.025	
3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	.010	
3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61	.005	
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57	.100	22
2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	.050	
2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00	.025	
3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	.010	
3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55	.005	
1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55	.100	23
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76	.050	
2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97	.025	
3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	.010	
3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48	.005	
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53	.100	24
2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73	.050	
2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94	.025	
3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	.010	
3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43	.005	
1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52	.100	25
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71	.050	
2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91	.025	
3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	.010	
3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38	.005	
1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50	.100	26
2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69	.050	
2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88	.025	
3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	.010	
3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33	.005	
1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	.100	27
2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67	.050	
2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	.025	
3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	.010	
3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29	.005	
1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	.100	28
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	.050	
2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	.025	
3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	.010	
3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25	.005	

Denominator *df*	Probability of a larger *F*	Numerator *df*								
		1	2	3	4	5	6	7	8	9
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01
120	.100	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	.050	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
	.010	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81
∞	.100	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63
	.050	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11
	.010	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62

Source: A portion of "Tables of percentage points of the inverted beta (*F*) distribution," *Biometrika*, vol. 33 (1943) by M. Merrington and C. M. Thompson and from Table 18 of *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, 1954, edited by E. S. Pearson and H. O. Hartley. Reproduced with permission of the authors, editors, and *Biometrika* trustees.

Numerator *df*											
10	12	15	20	24	30	40	60	120	∞	*P*	*df*
1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47	.100	29
2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64	.050	
2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81	.025	
3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03	.010	
3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21	.005	
1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46	.100	30
2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62	.050	
2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79	.025	
2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01	.010	
3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18	.005	
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38	.100	40
2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51	.050	
2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64	.025	
2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80	.010	
3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93	.005	
1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29	.100	60
1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39	.050	
2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48	.025	
2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60	.010	
2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69	.005	
1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19	.100	120
1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25	.050	
2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31	.025	
2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38	.010	
2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43	.005	
1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00	.100	∞
1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00	.050	
2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00	.025	
2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00	.010	
2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00	.005	

Table A-5. Upper 5% points, Q (Nash, 1965).

	Number of treatments, a								
Degrees of freedom	2	3	4	5	6	7	8	9	10
1	18.0	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99
3	4.50	5.91	6.82	7.50	8.04	8 48	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.44	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47

Table 9.3 is taken from J. Parchares "Table of the Upper 10% Pcints of the Studentized Range", Biometrika 46: 461-466 (1959) with additional table by Dr. Leon Harter by permission of Professor E. S. Pearson, Editor of Biometrika.

☆U.S. GOVERNMENT PRINTING OFFICE:1983—776-976 / 66